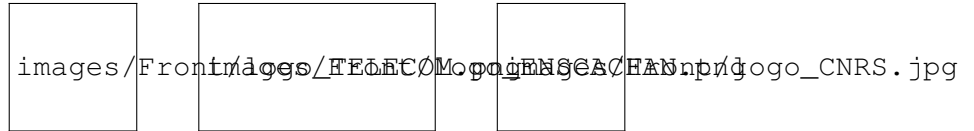


École Doctorale ED130 “Informatique, télécommunications et électronique de Paris”



Machine Learning and Extremes for Anomaly Detection

—

Apprentissage Automatique et Extrêmes pour la Détection d’Anomalies

Thèse pour obtenir le grade de docteur délivré par

TELECOM PARISTECH

Spécialité “Signal et Images”

présentée et soutenue publiquement par

Nicolas GOIX

le 1er Octobre 2016

LTCL, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France

Jury :

Gérard Biau	Professeur, Université Pierre et Marie Curie	Examineur
Gilles Blanchard	Professeur, Universität Potsdam	Rapporteur
Stéphane Boucheron	Professeur, Université Paris Diderot	Examineur
Stéphan Cléménçon	Professeur, Télécom ParisTech	Directeur
Alexandre Gramfort	Maitre de Conférence, Télécom ParisTech	Examineur
Anne Sabourin	Maitre de Conférence, Télécom ParisTech	Co-directeur
Johan Segers	Professeur, Université catholique de Louvain	Rapporteur
Jean-Philippe Vert	Professeur, Mines ParisTech	Examineur

List of Contributions

Journal

- Sparse Representation of Multivariate Extremes with Applications to Anomaly Detection. (Under review for Journal of Multivariate Analysis).
Authors: Goix, Sabourin, and Cléménçon.

Conferences

- On Anomaly Ranking and Excess-Mass Curves. (AISTAT 2015).
Authors: Goix, Sabourin, and Cléménçon.
- Learning the dependence structure of rare events: a non-asymptotic study. (COLT 2015).
Authors: Goix, Sabourin, and Cléménçon.
- Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. (AISTAT 2016).
Authors: Goix, Sabourin, and Cléménçon.
- How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms? (submitted to NIPS 2016).
Authors: Goix and Thomas.
- One-Class Splitting Criteria for Random Forests with Application to Anomaly Detection. (submitted to NIPS 2016).
Authors: Goix, Brault, Drougard and Chiapino.

Workshops

- Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. (NIPS 2015 Workshop on Nonparametric Methods for Large Scale Representation Learning).
Authors: Goix, Sabourin, and Cléménçon.

- How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms? (ICML 2016, Workshop on Anomaly Detection). Co-winner of the Best Paper Award, sponsored by Google.
Author: Goix.

Scikit-Learn implementations

- Isolation Forest: <https://github.com/scikit-learn/scikit-learn/pull/4163>
Authors: Goix and Gramfort
- Local Outlier Factor: <https://github.com/scikit-learn/scikit-learn/pull/5279>
Authors: Goix and Gramfort

Remerciements

Contents

List of Contributions	iii
List of Figures	xi
List of Tables	xiii
1 Summary	1
1.1 Introduction	1
1.2 Anomaly Detection and Scoring Functions	2
1.3 M-estimation and Scoring function Criterion	4
1.3.1 Minimum Volume sets	5
1.3.2 Mass-Volume curve	6
1.3.3 The Excess-Mass criterion	7
1.4 Evaluation of AD algorithms	9
1.5 One-Class Random Forests	11
1.6 Extreme Values Analysis through STDF estimation	14
1.7 Sparse Representation of Multivariate Extremes	16
1.7.1 Context: multivariate extreme values in large dimension	16
1.8 Scikit-learn contributions	19
1.9 Conclusion and Scientific Output	19
I Preliminaries	23
2 Background on Anomaly Detection through Scikit-Learn	25
2.1 What is Anomaly Detection?	25
2.2 Three efficient Anomaly Detection Algorithms	27
2.2.1 What is Scikit-learn?	27
2.2.2 One-class SVM	28
2.2.3 Local Outlier Factor algorithm	31
2.2.4 Isolation Forest	33
3 Concentration Inequalities from the Method of bounded differences	39
3.1 Two fundamental results	39
3.1.1 Preliminary definitions	39
3.1.2 Inequality for Bounded Random Variables	40
3.1.3 Bernstein-type Inequality (with variance term)	42
3.2 Famous Inequalities	43
3.3 Links with Statistical Learning and VC theory	46

3.4	Sharper VC-bounds through a Bernstein-type inequality	48
4	Extreme Value Theory	53
4.1	Univariate Extreme Value Theory	54
4.2	Extension to the Multivariate framework	56
II	An Excess-Mass based Performance Criterion	61
5	On Anomaly Ranking and Excess-Mass Curves	63
5.1	Introduction	63
5.2	Background and related work	65
5.3	The Excess-Mass curve	67
5.4	A general approach to learn a scoring function	70
5.5	Extensions - Further results	73
5.5.1	Distributions with non compact support	73
5.5.2	Bias analysis	75
5.6	Simulation examples	77
5.7	Conclusion	77
5.8	Illustrations	79
5.9	Detailed Proofs	79
6	How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?	87
III	One Class Random Forests	89
7	One Class Splitting Criteria for Random Forests with Application to Anomaly Detection	91
7.1	Introduction	91
IV	Accuracy on Extreme Regions	93
8	Learning the dependence structure of rare events: a non-asymptotic study	95
8.1	Introduction	95
8.2	Background on the stable tail dependence function	96
8.3	A VC-type inequality adapted to the study of low probability regions	97
8.4	A bound on the STDF	100
8.5	Discussion	106
8.6	Proof of Theorem 8.1	106
8.7	Note on Remark 8.5	109
9	Sparse Representation of Multivariate Extremes	111
9.1	Introduction	111
9.1.1	Context: multivariate extreme values in large dimension	111
9.1.2	Application to Anomaly Detection	114
9.2	Multivariate EVT Framework and Problem Statement	115
9.2.1	Statement of the Statistical Problem	116

9.2.2	Regularity Assumptions	119
9.3	A non-parametric estimator of the subcones' mass : definition and preliminary results	121
9.3.1	A natural empirical version of the exponent measure μ	121
9.3.2	Accounting for the non asymptotic nature of data: epsilon-thickening.	122
9.3.3	Preliminaries: uniform approximation over a VC-class of rectangles	123
9.3.4	Bounding empirical deviations over thickened rectangles	127
9.3.5	Bounding the bias induced by thickened rectangles	128
9.3.6	Main result	128
9.4	Application to Anomaly Detection	132
9.4.1	Background on anomaly detection	132
9.4.2	DAMEX Algorithm: Detecting Anomalies among Multivariate Extremes	133
9.5	Experimental results	136
9.5.1	Recovering the support of the dependence structure of generated data	136
9.5.2	Sparse structure of extremes (wave data)	137
9.5.3	Application to Anomaly Detection on real-world data sets	138
9.6	Conclusion	140
9.7	Technical proofs	143
9.7.1	Proof of Lemma 9.5	143
9.7.2	Proof of Lemma 9.6	144
9.7.3	Proof of Proposition 9.8	146
9.7.4	Proof of Lemma 9.10	150
9.7.5	Proof of Remark 9.14	152
9.8	Experiments curves	152
10	Conclusion & Perspectives	153
	Bibliography	155

List of Figures

1.1	Mass-Volume at level α	6
1.2	Truncated cones in 3D	18
1.3	Truncated ϵ -cones in 2D	18
2.1	LOF example	32
2.2	Anomalies are isolated more quickly	33
2.3	Convergence of the averaged depth	33
2.4	Isolation Forest example	34
2.5	gaussian normal data with one single mode	36
2.6	gaussian normal data with two modes	36
2.7	gaussian normal data with two strongly separate modes	36
4.1	Extreme Value Distribution with $\alpha = 2$	55
5.1	EM curves depending on densities	68
5.2	Comparison between $MV^*(\alpha)$ and $EM^*(t)$	68
5.3	Unsuccessful mass-volume criterion optimization	74
5.4	Optimal and realized EM curves	78
5.5	Zoom near 0	78
5.6	EM_G for different l	78
5.7	density and scoring functions	79
9.1	Truncated cones in 3D	117
9.2	Truncated ϵ -rectangles in 2D	117
9.3	Estimation procedure	123
9.4	Level sets of s_n on simulated 2D data	135
9.5	sub-cone dimensions of wave data	138
9.6	SF dataset, default parameters	141
9.7	SF dataset, larger ϵ	142
9.8	SF dataset, larger ϵ	142
9.9	SA dataset, default parameters	143
9.10	forestcover dataset, default parameters	143
9.11	http dataset, default parameters	144

List of Tables

1.1	Summary of notation	3
2.1	SVM vs. OCSVM (hard-margin separation)	29
2.2	SVM vs. OCSVM (ν -soft margin separation)	30
9.1	Support recovering on simulated data	137
9.2	Total number of sub-cones of wave data	138
9.3	Datasets characteristics	140
9.4	Results on extreme regions with standard parameters $(k, \epsilon) = (n^{1/2}, 0.01)$. . .	140
9.5	Results on extreme regions with lower $\epsilon = 0.1$	141

1.1 Introduction

Anomaly, from grec $\alpha\nu\omega\mu\alpha\lambda\iota\alpha$, asperity, irregular, not the same (an-homalos), refers to a gap, a deviation with respect to some norm or basis, reflecting the expected behavior. We call *anomaly* the object inducing a gap, the observation which deviates from normality. In various fields, the following situation occurs: an expert aims at predicting a phenomenon based on previous observations. The most basic case is when one wants to predict some binary characteristic of some new observations/records, given previous ones. For instance one can think of a doctor aiming to predict if an incoming patient has some fixed pathology or not, using previous patients record (age, history, gender, blood pressure) associated with their true **label** (having the pathology or not). This case is an example of *binary classification*: the doctor aims to find a rule to **predict** the label of a new patient (the latter being characterized by its record). This rule is called a *classifier* and it has to be built, or *trained*, on previous records. Intuitively, the classifier will predict the same diagnosis for similar records, in a sense that should be learned precisely.

Two cases can be distinguished. If labels of previous patients are known/available, *i.e.* previous patients are known to be sick or healthy: the classification task is said to be **supervised**. If training labels are unknown, the classification is said **unsupervised**. Following our example, the doctor has to find two patterns (or clusters), healthy/sick, each containing similar patient records.

Anomaly detection occurs when one label is highly under-represented for training, for instance if very few patients have the pathology in the training database. Thus, **supervised anomaly detection** boils down to *rare class mining*, namely supervised classification on highly unbalanced classes. As to **unsupervised anomaly detection** (also simply called outlier detection), it generally assumes that the database has a hidden ‘normal’ model, and anomalies are observations which deviate from this model. The doctor wants to find records which deviate from the vast majority of those of his previous patients. His task is in some way simplified if he knows all of its previous patients to be healthy: it is easier for him to learn the ‘normal’ model, *i.e.* the typical record of a healthy patient, to be confronted with new records. This is the so-called **semi-supervised anomaly detection** framework (also called novelty detection), where the training database only contains normal instances.

This chapter is organized as follows. First in Section 1.2, the anomaly detection task is formally introduced as well as the concept of scoring function. Two criteria for ‘being a good scoring function’ are presented in Section 1.3, allowing an M-estimation approach. In Section 1.6, we introduce the extreme value theory (EVT) and the *stable tail dependence function* (STDF), to estimate the dependence structure of rare events. Section 1.7 shows that multivariate EVT can be useful to produce scoring functions accurate on low probability regions.

Notations Throughout this document, \mathbb{N} denotes the set of natural numbers while \mathbb{R} and \mathbb{R}_+ respectively denote the sets of real numbers and nonnegative real numbers. Arbitrary sets are denoted by calligraphic letters such as \mathcal{G} , and $|\mathcal{G}|$ stands for the number of elements in \mathcal{G} . We denote vectors by bold lower case letters. For a vector $\mathbf{x} \in \mathbb{R}^d$ and $i \in \{1, \dots, d\}$, x_i denotes the i^{th} component of \mathbf{x} . The inner product between two vectors is denoted by $\langle \cdot, \cdot \rangle$. $\|\cdot\|$ denotes an arbitrary (vector or matrix) norm and $\|\cdot\|_p$ the L_p norm. Throughout this thesis, $\mathbb{P}[A]$ denotes the probability of the event $A \in \Omega$, the underlying probability space being $(\Omega, \mathcal{F}, \mathbb{P})$. We denote by $\mathbb{E}[X]$ the expectation of the random variable X . $X \stackrel{d}{=} Y$ means that X and Y are equal in distribution and $X_n \xrightarrow{d} Y$ means that (X_n) converges to Y in distribution. We often use the abbreviation $\mathbf{X}_{1:n}$ to denote an *i.i.d.* sample $(\mathbf{X}_1, \dots, \mathbf{X}_n)$. A summary of the notations is given in Table 1.1.

1.2 Anomaly Detection and Scoring Functions

From a probabilistic point of view, there are different way of modeling normal and abnormal behaviors, which leads to different methodologies. One natural probabilistic model is to assume two different generating processes for normal and abnormal data. Normal data (resp. abnormal data) are generated according to some distribution F (resp. G). The general underlying distribution is then a mixture of F and G . The goal is to find out whether a new observation \mathbf{x} has been generated from F , or from G . The optimal way to resolve theoretically this problem is the likelihood ratio test, also called Neyman-Pearson test. If $(dF/dG)(\mathbf{x}) > t$ with some $t > 0$ threshold, then \mathbf{x} has been drawn from F . Otherwise, \mathbf{x} has been drawn from G . This boils down to estimating the *density level set* $\{\mathbf{x}, (dF/dG)(\mathbf{x}) > t\}$. As anomalies are very rare, their structure cannot be observed in the data, in particular their distribution G . It is common and convenient to replace G in the problem above by the Lebesgue measure, so that it boils down to estimating density level set of F . This comes back to assume that anomalies are uniformly distributed on the support of the normal distribution (Blanchard et al., 2010). This assumption is thus implicitly made by a majority of works on novelty detection. We observe data in \mathbb{R}^d from the normal class only, with an underlying distribution F and with a density $f : \mathbb{R}^d \rightarrow \mathbb{R}$. The goal is to identify characteristics of this normal class, such as its support $\{\mathbf{x}, f(\mathbf{x}) > 0\}$ or some density level set $\{\mathbf{x}, f(\mathbf{x}) > t\}$ with $t > 0$ close to 0.

Notation	Description
$c.d.f.$	cumulative distribution function
$r.v.$	random variable
\mathbb{R}	Set of real numbers
\mathbb{R}_+	Set of nonnegative real numbers
\mathbb{R}^d	Set of d -dimensional real-valued vectors
$\text{Leb}(\cdot)$	Lebesgue measure on \mathbb{R} or \mathbb{R}^d
$(\cdot)_+$	positive part
\vee	maximum operator
\wedge	minimum operator
\mathbb{N}	Set of natural numbers, i.e., $\{0, 1, \dots\}$
\mathcal{G}	An arbitrary set
$ \mathcal{G} $	Number of elements in \mathcal{G}
\mathbf{x}	An arbitrary vector
$\mathbf{x} < \mathbf{y}$	component-wise vector comparison
\mathbf{m} (for $m \in \mathbb{R}$)	vector (m, \dots, m)
$\mathbf{x} < m$	means $\mathbf{x} < \mathbf{m}$
x_j	The j^{th} component of \mathbf{x}
$\delta_{\mathbf{a}}$	Dirac mass at point $\mathbf{a} \in \mathbb{R}^d$
$\lfloor \cdot \rfloor$	integer part
$\langle \cdot, \cdot \rangle$	Inner product between vectors
$\ \cdot \ $	An arbitrary norm
$\ \cdot \ _p$	L_p norm
$A \Delta B$	symmetric difference between sets A and B
$(\Omega, \mathcal{F}, \mathbb{P})$	Underlying probability space
\mathcal{S}	functions $s : \mathbb{R}^d \rightarrow \mathbb{R}_+$ integrable w.r.t. Lebesgue measure (scoring functions)
\xrightarrow{d}	Weak convergence of probability measures or r.v.
\mathbf{X}	A r.v. with values in \mathbb{R}^d
$\mathbb{1}_{\mathcal{E}}$	indicator function event \mathcal{E}
$Y_{(1)} \leq \dots \leq Y_{(n)}$	order statistics of Y_1, \dots, Y_n
$\mathbf{X}_{1:n}$	An <i>i.i.d.</i> sample $(\mathbf{X}_1, \dots, \mathbf{X}_n)$
$\mathbb{P}[\cdot]$	Probability of event
$\mathbb{E}[\cdot]$	Expectation of random variable
$\text{Var}[\cdot]$	Variance of random variable

TABLE 1.1: Summary of notation.

This *one-class classification* problem is different than *distinguishing* between several classes as done in standard classification. Also, unsupervised anomaly detection is often viewed as a one-class classification problem, where training data are polluted by a few elements of the abnormal class: it appeals for one-class algorithms *robust to anomalies*.

A natural idea for estimating density level sets is to compute an estimate of the density and to consider the associated plug-in density level set. The density is generally estimated using non-parametric kernel estimator or maximum likelihood estimator from some parametric family of functions. But these methods does not scale well with the dimension. Such methods try somehow to capture more information than needed for our level set estimation task, such as local properties of the density which are useless here. Indeed, it turns out that for any increasing transform T , the level sets of $T \circ f$ are exactly those of f . Thus, it suffices to estimate any representant of the class of all increasing transforms of f , to obtain density level sets estimates.

Intuitively, it is enough to estimate the preorder (the *scoring*) induced by f on \mathbb{R}^d . Let us define a *scoring function* as any measurable function $s : \mathbb{R}^d \rightarrow \mathbb{R}_+$ integrable w.r.t. the Lebesgue measure $\text{Leb}(\cdot)$, and \mathcal{S} the space of all scoring functions. Any scoring function defines a preorder on \mathbb{R}^d and thus a ranking on a set of new observations. This ranking can be interpreted as a degree of abnormality, the lower $s(x)$, the more abnormal x . Note incidentally that most anomaly detection algorithms return more than a binary label, normal/abnormal. They compute first a scoring function, which is converted to a binary prediction, typically by imposing some threshold based on its statistical distribution.

Suppose we are interested in learning a scoring function s whose induced preorder is ‘close’ to that of f , or equivalently whose induced level sets are close to those of f . The problem is to turn this notion of proximity into a criterion \mathcal{C} , optimal scoring functions s^* being then defined as those optimizing \mathcal{C} . In the density estimation framework, the uniform difference $\|f - \hat{f}\|_\infty$ is a common criterion to assess the quality of the estimation. We would like a similar criterion but which is invariant by increasing transformation of the output \hat{f} . In other words, the criterion should be defined such that the collection of level sets of an optimal scoring function $s^*(x)$ coincides with that related to f , and any increasing transform of the density should be optimal regarding \mathcal{C} . More formally, we are going to consider $\mathcal{C}^\Phi(s) = \|\Phi(s) - \Phi(f)\|$ (instead of $\|s - f\|$) with $\Phi : \mathbb{R} \rightarrow \mathbb{R}_+$ verifying $\Phi(T \circ s) = \Phi(s)$ for any scoring function s and increasing transform T . Here $\Phi(s)$ denotes either the mass-volume curve MV_s of s or its excess-mass curve EM_s , which are defined in the next section.

This criterion which measures the quality of a scoring function is then a tool for building/learning a good scoring function. According to the Empirical Risk Minimization (ERM) paradigm, a scoring function is built by optimizing an empirical version $\mathcal{C}_n(s)$ of the criterion over an adequate set of scoring functions \mathcal{S}_0 of controlled complexity (e.g. a class of finite VC dimension).

The next section describes two criteria, which are functional due to the global nature of the problem just like the *Receiver Operating Characteristic* (ROC) and *Precision-Recall* (PR) curves, and which are admissible with respect to the requirements listed above. These functional criteria extend somehow the concept of ROC curve to the unsupervised setup.

1.3 M-estimation and Scoring function Criterion

This section is a summary of Chapter 5, which is based on previous work published in Goix et al. (2015c). We provide a brief overview of the mass-volume curve criterion introduced in Cl  men  on & Jakubowicz (2013), which is based on the notion of minimum volume sets. We then exhibit the main drawbacks of this approach, and propose an alternative criterion, the excess-mass curve to circumscribe these drawbacks.

1.3.1 Minimum Volume sets

The notion of minimum volume set (Polonik (1997); Einmahl & Mason (1992)) has been introduced to describe regions where a multivariate $r.v.$ $\mathbf{X} \in \mathbb{R}^d$ takes its values with highest/smallest probability. Let $\alpha \in (0, 1)$, a minimum volume set Γ_α^* of mass at least α is any solution of the constrained minimization problem

$$\min_{\Gamma \text{ borelian}} \text{Leb}(\Gamma) \text{ subject to } \mathbb{P}(\mathbf{X} \in \Gamma) \geq \alpha, \quad (1.1)$$

where the minimum is taken over all measurable subsets Γ of \mathbb{R}^d . It can be shown that every density level set is a minimum volume set for a specific mass target, and that the reverse is true if the density has no flat part. In the remainder of this section we suppose that F has a density $f(x)$ w.r.t. the Lebesgue measure on \mathbb{R}^d satisfying the following assumptions:

A₁ *The density f is bounded.*

A₂ *The density f has no flat parts: $\forall c \geq 0, \mathbb{P}\{f(\mathbf{X}) = c\} = 0$.*

Under the hypotheses above, for any $\alpha \in (0, 1)$, there exists a unique minimum volume set Γ_α^* , whose mass is equal to α exactly. The (generalized) quantile function is then defined by:

$$\forall \alpha \in (0, 1), \quad \lambda^*(\alpha) := \text{Leb}(\Gamma_\alpha^*).$$

Additionally, the mapping λ^* is continuous on $(0, 1)$ and uniformly continuous on $[0, 1 - \epsilon]$ for all $\epsilon \in (0, 1)$ – when the support of F is compact, uniform continuity holds on the whole interval $[0, 1]$.

Estimates $\hat{\Gamma}_\alpha^*$ of minimum volume sets are built by replacing the unknown probability distribution F by its empirical version $F_n = (1/n) \sum_{i=1}^n \delta_{\mathbf{X}_i}$ and restricting optimization to a collection \mathcal{A} of borelian subsets of \mathbb{R}^d . \mathcal{A} is assumed to be rich enough to include all density level sets, or at least reasonable approximates of the latter. In Polonik (1997), limit results are derived for the generalized empirical quantile process $\{\text{Leb}(\hat{\Gamma}_\alpha^*) - \lambda^*(\alpha)\}$ (under the assumption in particular that \mathcal{A} is a Glivenko-Cantelli class for F). In Scott & Nowak (2006), it is proposed to replace the level α by $\alpha - \phi_n$ where ϕ_n plays the role of tolerance parameter (of the same order as the supremum $\sup_{\Gamma \in \mathcal{A}} |F_n(\Gamma) - F(\Gamma)|$), the complexity of the class \mathcal{A} being controlled by the VC dimension, so as to establish rate bounds. The statistical version of the Minimum Volume set problem then becomes

$$\min_{\Gamma \in \mathcal{A}} \text{Leb}(\Gamma) \text{ subject to } F_n(\Gamma) \geq \alpha - \phi_n.$$

The ensemble \mathcal{A} of borelian subsets of \mathbb{R}^d ideally offers both statistical and computational advantages; allowing for fast search as well as being sufficiently complex to capture the geometry of target density level sets – *i.e.* the ‘model bias’ $\inf_{\Gamma \in \mathcal{A}} \text{Leb}(\Gamma \Delta \Gamma_\alpha^*)$ should be small.

1.3.2 Mass-Volume curve

Let $s \in \mathcal{S}$ a scoring function. As defined in Cl  men  on & Jakubowicz (2013); Cl  men  on & Robbiano (2014), the mass-volume curve of s is the plot of the mapping

$$MV_s : \alpha \in (0, 1) \mapsto MV_s(\alpha) = \lambda_s \circ \alpha_s^{-1}(\alpha),$$

where H^{-1} denotes the pseudo-inverse of any cdf $H : \mathbb{R} \rightarrow (0, 1)$ and where α_s and λ_s are defined by

$$\begin{aligned} \alpha_s(t) &:= \mathbb{P}(s(\mathbf{X}) \geq t), \\ \lambda_s(t) &:= \text{Leb}(\{\mathbf{x} \in \mathbb{R}^d, s(\mathbf{x}) \geq t\}). \end{aligned} \quad (1.2)$$

This induces a partial ordering on the set of all scoring functions, in the sense that s is preferred to s' if $MV_s(\alpha) \leq MV_{s'}(\alpha)$ for all $\alpha \in (0, 1)$. Also, the mass-volume curve remains unchanged when applying any increasing transformation on s . It can be proven that $MV^*(\alpha) \leq MV_s(\alpha)$ for all $\alpha \in (0, 1)$ and any scoring function s , where $MV^*(\alpha)$ is the optimal value of the constrained minimization problem (1.1), namely

$$MV^*(\alpha) = \text{Leb}(\Gamma_\alpha^*) = \min_{\Gamma \text{ mes.}} \text{Leb}(\Gamma) \text{ subject to } \mathbb{P}(\mathbf{X} \in \Gamma) \geq \alpha. \quad (1.3)$$

Under assumptions **A**₁ and **A**₂, one may show that the curve MV^* is actually a MV curve, that is related to (any increasing transform of) the density f namely: $MV^* = MV_f$. The objective is then to build a scoring function \hat{s} depending on training data $\mathbf{X}_1, \dots, \mathbf{X}_n$ such that $MV_{\hat{s}}$ is (nearly) minimum everywhere, *i.e.* minimizing $\|MV_{\hat{s}} - MV^*\|_\infty := \sup_{\alpha \in [0,1]} |MV_{\hat{s}}(\alpha) - MV^*(\alpha)|$.

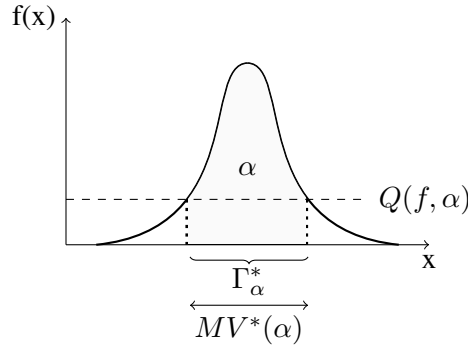


FIGURE 1.1: Mass-Volume at level α

The way of doing it consists in preliminarily estimating a collection of minimum volume sets related to target masses $0 < \alpha_1 < \dots < \alpha_K < 1$ forming a subdivision of $(0, 1)$ based on training data so as to define $s = \sum_k \mathbb{1}_{\{x \in \Gamma_{\alpha_k}^*\}}$. The analysis is done under adequate assumptions (related to \mathcal{G} , the perimeter of the $\Gamma_{\alpha_k}^*$'s and the subdivision step in particular) and for an appropriate choice of $K = K_n$. However, by construction, learning rate bounds are rather slow (of the order $n^{-1/4}$ namely) and cannot be established in the unbounded support situation.

But the four main drawbacks of this mass-volume curve criterion are the following.

- 1) When used as a performance criterion, the Lebesgue measure of possibly very complex sets has to be computed.
- 2) When used as a performance criterion, there is no simple manner to compare MV-curves since the area under the curve is potentially infinite.
- 3) When used as a learning criterion (in the ERM paradigm), it produces level sets which are not necessarily nested, and then inaccurate scoring functions.
- 4) When used as a learning criterion, the learning rates are rather slow (of the order $n^{-1/4}$ namely), and cannot be established in the unbounded support situation.

In the following section, and as a contribution of this thesis, an alternative functional criterion is proposed, obtained by exchanging objective and constraint functions in (1.1). The drawbacks of the mass-volume curve criterion are resolved excepting the first one, and it is shown that optimization of an empirical discretized version of this performance measure yields scoring rules with convergence rates of the order $\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})$. In addition, the results can be extended to the situation where the support of the distribution F is not compact. Also, when relaxing the assumption made in the mass-volume curve analysis that all level sets are included in our minimization class \mathcal{A} , a control of the model bias is established. Last but not least, we derive (non-statistical) theoretical properties verified by this criterion, which corroborate its adequacy as a metric on preorders/level sets summarized in scoring functions.

1.3.3 The Excess-Mass criterion

We propose an alternative performance criterion which relies on the notion of *excess mass* and *density contour clusters*, as introduced in the seminal contribution Polonik (1995). The main idea is to consider a Lagrangian formulation of a constrained minimization problem, obtained by exchanging constraint and objective in (1.1): for $t > 0$,

$$\max_{\Omega \text{ borelian}} \{ \mathbb{P}(\mathbf{X} \in \Omega) - t \text{Leb}(\Omega) \}. \quad (1.4)$$

We denote by Ω_t^* any solution of this problem. This formulation offers certain computational and theoretical advantages both at the same time: when letting (a discretized version of) the Lagrangian multiplier t increase from 0 to infinity, one may easily obtain solutions of empirical counterparts of (1.4) forming a *nested* sequence of subsets of the feature space, avoiding thus deteriorating rate bounds by transforming the empirical solutions so as to force monotonicity. The **optimal Excess-Mass curve** related to a given probability distribution F is defined as the plot of the mapping

$$t > 0 \mapsto EM^*(t) := \max_{\Omega \text{ borelian}} \{ \mathbb{P}(\mathbf{X} \in \Omega) - t \text{Leb}(\Omega) \}.$$

Equipped with the notation above, we have: $EM^*(t) = \mathbb{P}(\mathbf{X} \in \Omega_t^*) - t\text{Leb}(\Omega_t^*)$ for all $t > 0$. Notice also that $EM^*(t) = 0$ for any $t > \|f\|_\infty := \sup_{x \in \mathbb{R}^d} |f(x)|$. The **Excess-Mass curve** of $s \in \mathcal{S}$ w.r.t. the probability distribution F of a random variable \mathbf{X} is the plot of the mapping

$$EM_s : t \in [0, \infty[\mapsto \sup_{A \in \{(\Omega_{s,t})_{t>0}\}} \{\mathbb{P}(\mathbf{X} \in A) - t\text{Leb}(A)\}, \quad (1.5)$$

where $\Omega_{s,t} = \{x \in \mathbb{R}^d, s(x) \geq t\}$ for all $t > 0$. One may also write EM_s in terms of λ_s and α_s defined in (1.2), $EM_s(t) = \sup_{u>0} \alpha_s(u) - t\lambda_s(u)$. Finally, under assumption **A₁**, we have $EM_s(t) = 0$ for every $t > \|f\|_\infty$.

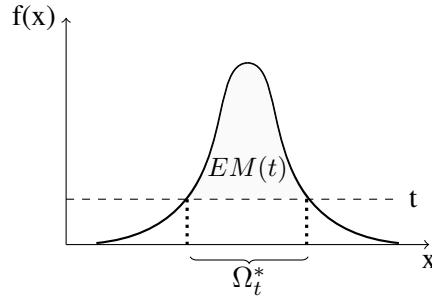


Figure 2: Excess-Mass curve

Maximizing EM_s can be viewed as recovering a collection of subsets $(\Omega_t^*)_{t>0}$ with maximum mass when penalized by their volume in a linear fashion. An optimal scoring function is then any $s \in \mathcal{S}$ with the Ω_t^* 's as level sets, for instance any scoring function of the form

$$s(x) = \int_{t=0}^{+\infty} \mathbb{1}_{x \in \Omega_t^*} a(t) dt,$$

with $a(t) > 0$ (observe that $s(x) = f(x)$ for $a \equiv 1$). The mapping EM_s is non increasing on $(0, +\infty)$, takes its values in $[0, 1]$ and satisfies, $EM_s(t) \leq EM^*(t)$ for all $t \geq 0$. In addition, for $t \geq 0$ and any $\epsilon > 0$, we have

$$\inf_{u>0} \epsilon \text{Leb}(\{s > u\} \Delta_\epsilon \{f > t\}) \leq EM^*(t) - EM_s(t) \leq \|f\|_\infty \inf_{u>0} \text{Leb}(\{s > u\} \Delta \{f > t\})$$

with $\{s > u\} \Delta_\epsilon \{f > t\} := \{f > t + \epsilon\} \setminus \{s > u\} \sqcup \{s > u\} \setminus \{f > t - \epsilon\}$. Thus the quantity $EM^*(t) - EM_s(t)$ measures how well level sets of s can approximate those of the underlying density. Under some reasonable conditions (see Goix et al. (2015c), Prop.1), we also have for $\epsilon > 0$,

$$\sup_{t \in [\epsilon, \|f\|_\infty]} |EM^*(t) - EM_s(t)| \leq C \inf_{T \in \mathcal{T}} \|f - T \circ s\|_\infty$$

where the infimum is taken over the set \mathcal{T} of all measurable increasing transforms $T : \mathbb{R}_+ \rightarrow \mathbb{R}_+$. The previous inequalities reveal that $\|EM^* - EM_s\|_\infty$ can be interpreted as a pseudo distance either between the level sets of s and those of the true density f , or between the preorders induced by s and f .

The concept of EM-curve provides a simple way to compare scoring functions but optimizing such a functional criterion is far from straightforward. As proposed in Cl  men  on & Jakubowicz (2013) for the MV criterion, optimization is done over some representative class of scoring functions, hopefully rich enough to provide a good approximation (small model bias) while simple enough to control the convergence rate. Here we consider scoring functions of the form

$$s_N(x) := \sum_{k=1}^N a_k \mathbb{1}_{x \in \hat{\Omega}_{t_k}}, \quad \text{with } \hat{\Omega}_{t_k} \in \mathcal{G}$$

where \mathcal{G} is a VC-class of subset of \mathbb{R}^d . We arbitrary take $a_k := (t_k - t_{k+1})$ so that the $\hat{\Omega}_{t_k}$'s correspond exactly to t_k level sets $\{s \geq t_k\}$. Then, maximizing the Excess-Mass functionnal criterion is done by sequentially resolving, for $k = 1, \dots, N$,

$$\hat{\Omega}_{t_k} \in \arg \max_{\Omega \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \in \Omega} - t_k \text{Leb}(\Omega).$$

The $\hat{\Omega}_{t_k}$'s solution of these optimization problems can always be chosen in such a way that they are nested (unlike the analogous optimization problem for the mass-volume criterion). In other words, an inclusion constraint can be incorporated into the previous optimization problem, without affecting the quality of the solution picked up. It allows to avoid forcing the solutions to be nested, yielding stronger convergence rates. In the mass-volume criterion M-estimation framework, assumptions are made stipulating that the support of the distribution is compact, and that the VC-class \mathcal{G} contains the true density level sets. Here we relax these assumptions, the first one by choosing adaptive levels t_k , and the second one by deriving a bias study. This is detailed in Chapter 5.

1.4 Evaluation of AD algorithms

This is a summary of Chapter 6, which is based on a workshop paper (Goix, 2016) and a submitted work (Goix & Thomas, 2016).

When sufficient labeled data are available, classical criteria based on ROC (Provost et al., 1997, 1998; Fawcett, 2006) or PR (Davis & Goadrich, 2006; Cl  men  on & Vayatis, 2009) curves can be used to compare the performance of unsupervised anomaly detection algorithms. However, in many situations, few or no data are labeled. This calls for alternative criteria one can compute on non-labeled data.

While excess-mass and mass-volume curves quantities have originally been introduced to build scoring functions *via* Empirical Risk Minimization (ERM), the MV-curve has been used recently for the calibration of the One-Class SVM (Thomas et al., 2015). When used to attest the quality of some scoring function, the volumes induced become unknown and must be estimated, which

is challenging in large dimension if no prior knowledge on the form of these level sets is available. Besides, the accuracy of EM or MV curves as evaluation criteria has not been studied yet. Summarized in this section and as a contribution of this thesis, numerical performance scores based on EM and MV criteria (that do not require labels) are empirically shown to discriminate accurately (*w.r.t.* ROC or PR based criteria) between algorithms. A methodology based on feature sub-sampling and aggregating is also described and tested. This extends the use of these criteria to high-dimensional datasets and solves major drawbacks inherent to standard EM and MV curves.

Recall that the MV and EM curves of a scoring function s can be written as

$$MV_s(\alpha) = \inf_{u \geq 0} \text{Leb}(s \geq u) \text{ s.t. } \mathbb{P}(s(\mathbf{X}) \geq u) \geq \alpha \quad (1.6)$$

$$EM_s(t) = \sup_{u \geq 0} \mathbb{P}(s(\mathbf{X}) \geq u) - t \text{Leb}(s \geq u) \quad (1.7)$$

for any $\alpha \in (0, 1)$ and $t > 0$. The optimal curves are $MV^* = MV_f = MV_{T \circ f}$ and $EM^* = EM_f = EM_{T \circ f}$ for any increasing transform $T : \text{Im}(f) \rightarrow \mathbb{R}$. As curves cannot be trivially compared, consider the L^1 -norm $\|\cdot\|_{L^1(I)}$ with $I \subset \mathbb{R}$ an interval. As $MV^* = MV_f$ is below MV_s pointwise, $\arg \min_s \|MV_s - MV^*\|_{L^1(I)} = \arg \min \|MV_s\|_{L^1(I)}$. We thus define $\mathcal{C}^{MV}(s) = \|MV_s\|_{L^1(I^{MV})}$, which is equivalent to consider $\|MV_s - MV^*\|_{L^1(I^{MV})}$ as mentioned in the introduction. As we are interested in evaluating accuracy on large density level-sets, one natural interval I^{MV} would be for instance $[0.9, 1]$. However, MV diverges in 1 when the support is infinite, so that we arbitrarily take $I^{MV} = [0.9, 0.999]$. The smaller is $\mathcal{C}^{MV}(s)$, the better is the scoring function s . Similarly, we consider $\mathcal{C}^{EM}(s) = \|EM_s\|_{L^1(I^{EM})}$, this time with $I^{EM} = [0, EM^{-1}(0.9)]$, where $EM_s^{-1}(0.9) := \inf\{t \geq 0, EM_s(t) \leq 0.9\}$, as $EM_s(0)$ is finite (equal to 1).

As the distribution F of the normal data is generally unknown, MV and EM curves must be estimated. Let $s \in \mathcal{S}$ and $\mathbf{X}_1, \dots, \mathbf{X}_n$ be an i.i.d. sample with common distribution F and set $\mathbb{P}_n(s \geq t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{s(\mathbf{X}_i) \geq t}$. The empirical MV and EM curves of s are then simply defined as empirical version of (1.6) and (1.7),

$$\widehat{MV}_s(\alpha) = \inf_{u \geq 0} \{\text{Leb}(s \geq u) \text{ s.t. } \mathbb{P}_n(s \geq u) \geq \alpha\} \quad (1.8)$$

$$\widehat{EM}_s(t) = \sup_{u \geq 0} \mathbb{P}_n(s \geq u) - t \text{Leb}(s \geq u) \quad (1.9)$$

Finally, we obtain the empirical EM and MV based performance criteria:

$$\widehat{\mathcal{C}}^{EM}(s) = \|\widehat{EM}_s\|_{L^1(I^{EM})} \quad I^{EM} = [0, \widehat{EM}^{-1}(0.9)], \quad (1.10)$$

$$\widehat{\mathcal{C}}^{MV}(s) = \|\widehat{MV}_s\|_{L^1(I^{MV})} \quad I^{MV} = [0.9, 0.999]. \quad (1.11)$$

The methodology to scale the use of the EM and MV criteria to large dimensional data consists in sub-sampling training *and* testing data along features, thanks to a parameter d' controlling the number of features randomly chosen for computing the (EM or MV) score. Replacement is done after each draw of features F_1, \dots, F_m . A partial score $\hat{\mathcal{C}}_k^{MV}$ (resp. $\hat{\mathcal{C}}_k^{EM}$) is computed for each draw F_k using (1.10) (resp. (1.11)). The final performance criteria are obtained by averaging these partial criteria along the different draws of features. This methodology is described in Algorithm 1.

Algorithm 1 High-dimensional EM/MV: evaluate AD algorithms on high-dimensional data

Inputs: AD algorithm \mathcal{A} , data set $X = (x_i^j)_{1 \leq i \leq n, 1 \leq j \leq d}$, feature sub-sampling size d' , number of draws m .
for $k = 1, \dots, m$ **do**
 randomly select a sub-group F_k of d' features
 compute the associated scoring function $\hat{s}_k = \mathcal{A}((x_i^j)_{1 \leq i \leq n, j \in F_k})$
 compute $\hat{\mathcal{C}}_k^{EM} = \|\widehat{EM}_{\hat{s}_k}\|_{L^1(I^{EM})}$ using (1.10) or $\hat{\mathcal{C}}_k^{MV} = \|\widehat{MV}_{\hat{s}_k}\|_{L^1(I^{MV})}$ using (1.11)
end for
Return performance criteria:

$$\hat{\mathcal{C}}_{high.dim}^{EM}(\mathcal{A}) = \frac{1}{m} \sum_{k=1}^m \hat{\mathcal{C}}_k^{EM} \quad (\text{idem for MV})$$

Low-dimensional and high-dimensional EM/MV are tested *w.r.t.* three classical AD algorithms. A wide range on real labeled datasets are used in the benchmark. Experiments show that when one algorithm has better performance than another on some fixed dataset, according to both ROC and PR AUCs, one can expect to recover it without using labels with an accuracy of 82% in the novelty detection framework, and 77% in the unsupervised framework.

1.5 One-Class Random Forests

This is a summary of Chapter 7, which is based on submitted work (Goix et al., 2016a).

Building accurate scoring functions by optimizing EM or MV criteria is very challenging in practice, as when building classifiers by optimizing the ROC curve (Cl  men  on & Vayatis (2010)) in the supervised framework. More work is needed for these methods to be efficient in practice, particularly for the choice of the class of sets on which the optimization is done. Indeed, this class is *hopefully rich enough to provide a good approximation while simple enough to control the convergence rate*. This compromise is hard to achieve, especially in high dimension when no prior knowledge on the shape of the level sets is available. In this section, we propose a heuristic approach to build scoring functions using Random Forests (RFs) (Breiman, 2001; Genuer et al., 2008; Biau & Scornet, 2016). More formally, we adapt RFs to the one-class classification framework by introducing one-class splitting criteria.

Standard RFs are estimators that fit a number of decision tree classifiers on different random sub-samples of the dataset. Each tree is built recursively, according to a splitting criterion based on some impurity measure of a node. The prediction is done by an average over each tree prediction. In classification the averaging is based on a majority vote. Few attempts to transfer the idea of RFs to one-class classification have already been made (Désir et al., 2012; Liu et al., 2008; Shi & Horvath, 2012). No algorithm structurally extends (without second class sampling and without alternative base estimators) RFs to one-class classification.

We introduce precisely such a methodology. It builds on a natural adaptation of two-class splitting criteria to the one-class setting, as well as an adaptation of the two-class majority vote. In addition, it turns out that the one-class model promoted here corresponds to the asymptotic behavior of an adaptive (with respect to the tree growing process) outliers generating methodology.

One-class Model with parameters (n, α) . We consider a random variable $X : \Omega \rightarrow \mathbb{R}^d$ w.r.t. a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The law of X depends on another r.v. $y \in \{0, 1\}$, verifying $\mathbb{P}(y = 1) = 1 - \mathbb{P}(y = 0) = \alpha$. We assume that conditionally on $y = 0$, X follows a law F , and conditionally on $y = 1$ a law G :

$$\begin{aligned} X \mid y = 0 &\sim F, & \mathbb{P}(y = 0) &= 1 - \alpha, \\ X \mid y = 1 &\sim G, & \mathbb{P}(y = 1) &= \alpha. \end{aligned}$$

We model the one-class framework as follows. Among the n *i.i.d.* observations, we only observe those with $y = 0$ (the normal behavior), namely N realizations of $(X \mid y = 0)$, where N is itself a realization of a r.v. \mathbf{N} of law $\mathbf{N} \sim \text{Bin}(n, (1 - \alpha))$, the binomial distribution with parameters (n, p) . As outliers are not observed, we classically assume that G follows a uniform distribution on the hyper-rectangle \mathcal{X} containing all the observations, so that G has a constant density $g(x) \equiv 1/\text{Leb}(\mathcal{X})$ on \mathcal{X} .

One-class empirical analogues of two-class impurity measures are then obtained by replacing the quantities relative to the normal behavior by their empirical versions. The quantities relative to the unobserved second class (abnormal behavior) are naturally expressed using the uniform distribution assumption.

In this way, our one-class impurity improvement function corresponds to the two-class one, where empirical second class quantities have been replaced by their expectation assuming a uniform distribution.

But it also induces a major problem: those expectations, which are proportional to the volume of the node at stake, become very small when going deeper in the tree. In the two-class framework, the corresponding problem is when the second class is highly under-represented in the neighborhood of the observations. As we assume the second class to be uniform on a hyper-rectangle containing all the observations, this fact was expected, especially in large dimension (curse of dimensionality). As the quantities relative to the second class are very close to zero, one observes that the impurity criterion becomes constant when the split varies, and then useless.

Adaptive approach. A solution is to chose adaptively (*w.r.t.* the volume of each node) the number αn , which can be interpreted as the number of (hidden) outliers. Recall that neither n nor α is observed in $\text{One-Class-Model}(n, \alpha)$ defined above.

The idea is to make $\alpha(t) \rightarrow 1, n(t) \rightarrow \infty$ when the volume of node t goes to zero. In other word, instead of considering one fixed general model $\text{One-Class-Model}(n, \alpha)$, we adapt it to each node t , considering $\text{One-Class-Model}(n(t), \alpha(t))$ *before searching the best split*. We still consider the N normal observations as a realization of this model. When growing the tree, using $\text{One-Class-Model}(n(t), \alpha(t))$ as the volume of t becomes close to zero allows to maintain a high expected proportion of outliers in the node to be split. Of course, constraints have to be made to ensure consistency between all these models. For instance, recalling that the number N of normal observations is a realization of \mathbf{N} following a Binomial distribution with parameters $(n, 1 - \alpha)$, a first natural constraint on $(n(t), \alpha(t))$ is

$$(1 - \alpha)n = (1 - \alpha(t)) \cdot n(t) \quad \text{for all } t, \quad (1.12)$$

so that the expectation of \mathbf{N} remains unchanged. Then the asymptotic model (when the volume of t goes to 0) consists in fact in assuming that the number N of normal data we observed is a realization of a Poisson distribution $\mathcal{P}((1 - \alpha)n)$, and that an infinite number of outliers have been hidden. In the two class framework, this corresponds to observing an infinite number of uniformly distributed outliers, breaking the curse of dimensionality (see Chapter 7 for details).

One-Class RF algorithm. Let us summarize the algorithm in its most generic version. It has 7 parameters: *max_samples*, *max_features_tree*, *max_features_node*, γ , *max_depth*, *n_trees*, s_k . Each tree is classically grown on a random subset of both the input samples and the input features (Ho, 1998; Panov & Džeroski, 2007). This random subset is a sub-sample of size *max_samples*, with *max_features_tree* variables chosen at random without replacement (replacement is only done after the tree is grown). The tree is built by minimizing a one-class version of the Gini criterion (Gini, 1912), obtained by replacing empirical quantities related to the (unobserved) second class by theoretical ones. These correspond to a weighted uniform distribution, the weight increasing when the volume of the node decreases, in order to avoid highly unbalanced classes (volume vs. observations). Indeed when their depth increases, the nodes tend to have smaller volumes while keeping as much (normal) observations as they can.

New nodes are built (by minimizing this criterion) until the maximal depth *max_depth* is achieved. Minimization is done as introduced in (Amit & Geman, 1997), by defining a large number *max_features_node* of geometric features and searching over a random selection of these for the best split at each node. The forest is composed of a number *n_trees* of trees. The predicted score of a point x is given by $s_k(x)$, which is either the stepwise density estimate (induced by the forest) around x , the local density of a typical cell containing x or the averaged depth of x among the forest. Chapter 7 formally defines the one-class splitting criteria and provides an extensive benchmark of state-of-the-art anomaly detection algorithms.

1.6 Extreme Values Analysis through STDF estimation

This section is a summary of Chapter 8, which is based on previous work published in Goix et al. (2015b).

Recall that scoring functions are built by approaching density level sets/minimum volume sets of the underlying ‘normal’ density. As mentioned previously, for an anomaly detection purpose, we are interested in being accurate on level sets corresponding to high quantiles, namely with level t close to 0 – equivalently being accurate on minimum volume sets with mass constraint α close to 1. In the univariate case, suppose we want to consider the $(1 - p)^{th}$ quantile of the distribution F of a random variable X , for a given exceedance probability p , that is $x_p = \inf\{x \in \mathbb{R}, \mathbb{P}(X > x) \leq p\}$. For moderate values of p , a natural empirical estimate is $x_{p,n} = \inf\{x \in \mathbb{R}, 1/n \sum_{i=1}^n \mathbb{1}_{X_i > x} \leq p\}$. However, if p is very small, the finite sample X_1, \dots, X_n contains insufficient information and $x_{p,n}$ becomes irrelevant. This problem transfers in the multivariate case to learning density level sets with very low level, or equivalently scoring functions inducing such level sets. Extreme value theory specially addresses such issues, in the one-dimensional as well as in the multi-dimensional setup.

Preliminaries. Extreme Value Theory (EVT) develops models for learning the unusual rather than the usual. These models are widely used in fields involving risk management like finance, insurance, telecommunication or environmental sciences. One major application of EVT is to provide a reasonable assessment of the probability of occurrence of rare events.

To illustrate this point, suppose we want to manage the risk of a portfolio containing d different assets, $\mathbf{X} = (X_1, \dots, X_d)$. A fairly general purpose is then to evaluate the probability of events of the kind $\{X_1 \geq x_1 \text{ or } \dots \text{ or } X_d \geq x_d\}$, for large multivariate thresholds $\mathbf{x} = (x_1, \dots, x_d)$. Under not too stringent conditions on the regularity of \mathbf{X} ’s distribution, EVT shows that for large enough thresholds,

$$\mathbb{P}\{X_1 \geq x_1 \text{ or } \dots \text{ or } X_d \geq x_d\} \simeq l(p_1, \dots, p_d),$$

where l is the *stable tail dependence function* (STDF) and the p_j ’s are the marginal exceedance probabilities, $p_j = \mathbb{P}(X_j \geq x_j)$. The functional l characterizes the *dependence* among extremes. The *joint* distribution (over large thresholds) can thus be recovered from the knowledge of the marginal distributions together with the STDF l . In practice, l can be learned from ‘moderately extreme’ data, typically the k ‘largest’ ones among a sample of size n , with $k \ll n$.

Recovering the p_j ’s can be done following a well paved way: in the univariate case, EVT essentially consists in modeling the distribution of the maxima (*resp.* the upper tail) as a generalized extreme value distribution, namely an element of the Gumbel, Fréchet or Weibull parametric families (*resp.* by a generalized Pareto distribution).

In contrast, in the multivariate case, there is no finite-dimensional parametrization of the dependence structure. The latter being characterized by the STDF, estimating this functional is one of the main issues in multivariate EVT. Asymptotic properties of the empirical STDF have been widely studied, see Huang (1992), Drees & Huang (1998), Embrechts et al. (2000) and de Haan & Ferreira (2006) for the bivariate case, and Qi (1997), Einmahl et al. (2012) for the general multivariate case under smoothness assumptions.

However, no bounds exist on the finite sample error. The contribution summarized in the next section and published in Goix et al. (2015b) derives such non-asymptotic bounds. Our results do not require any assumption other than the existence of the STDF.

Learning the dependence structure of rare events. Classical VC inequalities aim at bounding the deviation of empirical from theoretical quantities on relatively simple classes of sets, called VC classes. These classes typically cover the support of the underlying distribution. However, when dealing with rare events, it is of great interest to have such bounds on a class of sets which only covers a small probability region and thus contains (very) few observations. This yields sharper bounds, since only differences between very small quantities are involved. The starting point of this analysis is the following VC-inequality stated below and proved in Chapter 8.

Theorem 1.1. *Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. realizations of a r.v. \mathbf{X} , a VC-class \mathcal{A} with VC-dimension $V_{\mathcal{A}}$. Consider the class union $\mathbb{A} = \cup_{A \in \mathcal{A}} A$, and let $p = \mathbb{P}(\mathbf{X} \in \mathbb{A})$. Then there is an absolute constant C so that for all $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}[\mathbf{X} \in A] - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| \leq C \left[\sqrt{p} \sqrt{\frac{V_{\mathcal{A}}}{n} \log \frac{1}{\delta}} + \frac{1}{n} \log \frac{1}{\delta} \right].$$

The main idea is as follows. The empirical estimator of the STDF is based on the empirical measure of ‘extreme’ regions, which are hit only with low probability. It is thus enough to bound maximal deviations on such low probability regions. The key consists in choosing an adaptive VC class, which only covers the latter regions, and on the other hand, to derive VC-type inequalities that incorporate p , the probability of hitting the class at all. The bound we obtain on the finite sample error is then:

Theorem 1.2. *Let T be a positive number such that $T \geq \frac{7}{2}(\frac{\log d}{k} + 1)$, and δ such that $\delta \geq e^{-k}$. Then there is an absolute constant C such that for each $n > 0$, with probability at least $1 - \delta$:*

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

where l is the STDF, and by definition $l(\mathbf{x}) = \lim_{t \rightarrow 0} t^{-1} \tilde{F}(t\mathbf{x})$ such that the second term in the bound is a bias term.

In this section, we have introduced and studied, in a non-parametric setting, a particular functional characterizing the extreme dependence structure. One other convenient (nonparametric) characterization of extreme dependence in the framework of multivariate EVT is the *angular measure*, which provides direct information about the probable ‘directions’ of extremes, that is, the relative contribution of each feature/coordinate of the ‘largest’ observations. In many applications, it is more convenient to work with the angular measure itself. The latter gives more direct information on the dependence structure and is able to reflect structural simplifying properties (*e.g.* sparsity as detailed below) which would not appear in extreme value copulas or in the STDF which are integrated version of the angular measure. However, non-parametric modeling of the angular measure faces major difficulties, stemming from its potentially complex structure, especially in a high dimensional setting. Further, from a theoretical point of view, non-parametric estimation of the angular measure has only been studied in the two dimensional case, in Einmahl et al. (2001) and Einmahl & Segers (2009), in an asymptotic framework. As another contribution of this thesis, the section below summarizes a novel methodology aiming at exhibiting a sparsity pattern within the dependence structure of extremes.

1.7 Sparse Representation of Multivariate Extremes

This section is a summary of Chapter 9, which is based on previous work published in Goix et al. (2016b) and on its long version Goix et al. (2016c) under review.

EVT has been intensively used in anomaly detection in the one-dimensional situation, see for instance Roberts (1999), Roberts (2000), Clifton et al. (2011), Clifton et al. (2008), Lee & Roberts (2008). In the multivariate setup, however, there is –to the best of our knowledge– no anomaly detection method relying on *multivariate* EVT. Until now, the multidimensional case has only been tackled by means of extreme value statistics based on univariate EVT. The major reason is the difficulty to scale up existing multivariate EVT models with the dimensionality. In the present work we bridge the gap between the practice of anomaly detection and multivariate EVT by proposing a method which is able to learn a sparse ‘normal profile’ of multivariate extremes and, as such, may be implemented to improve the accuracy of any usual anomaly detection algorithm.

1.7.1 Context: multivariate extreme values in large dimension

Parametric or semi-parametric estimation of the structure of multivariate extremes is relatively well documented in the statistical literature, see *e.g.* Coles & Tawn (1991); Fougères et al. (2009); Cooley et al. (2010); Sabourin & Naveau (2012) and the references therein. However, restrictive structural assumptions have to be made, stipulating *e.g.* that only some pre-definite subgroups of components may be concomitantly extremes, or, on the contrary, that all of them must be. In addition, their practical use is restricted to moderate dimensional problems (say, $d \leq$

10), otherwise simplifying modeling choices are needed, as *e.g.* in Stephenson (2009)). Finally, uncertainty assessment concerning the output of these models is made under the hypothesis that the training set is ‘asymptotic’, in the sense that one assumes that, above a fixed high threshold, the data are exactly distributed according to the limit distribution of extremes. In other words, the modeling error is ignored.

Non-parametric estimation of the angular measure has only been treated in the two dimensional case, in Einmahl et al. (2001) and Einmahl & Segers (2009), in an asymptotic framework. Here we extend the nonasymptotic study on STDF estimation (previous section) to the angular measure of extremes, restricted to a well-chosen class of sets, corresponding to lower-dimensional regions of the space. The objective is to learn a representation of the angular measure, rough enough to control the variance in high dimension and accurate enough to gain information about the ‘probable directions’ of extremes, both at the same time. This yields a –first– non-parametric estimate of the angular measure in any dimension, restricted to a class of subcones, with a non asymptotic bound on the error. Note incidentally that this method can also be used as a pre-processing stage, for dimensionality reduction purpose, before proceeding with a parametric or semi-parametric estimation which could benefit from the structural information issued in the first step. Such applications are beyond the scope of this thesis.

The framework we develop is non-parametric and lies at the intersection of support estimation, density estimation and dimensionality reduction: it consists in learning the support of a distribution (from training data), that can be decomposed into subcones, hopefully each of low dimension and to which some mass is assigned, defining empirical versions of probability measures of specific extreme regions. It produces a scoring function defined and specially accurate on extreme regions, which can thus be exploited to detect anomalies among extremes. Due to its moderate complexity –of order $dn \log n$ – this algorithm is suitable for the treatment of real word large-scale learning problems, and experimental results reveal a significantly increased performance on extreme regions compared with standard anomaly detection approaches.

In a wide range of situations, one may expect the occurrence of two phenomena:

- 1-** 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 (‘small’ is relative to the total number of groups 2^d).
- 2-** 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 .

The main purpose of this work is to introduce a data-driven methodology for identifying such faces, so as to reduce the dimensionality of the problem and thus to learn a sparse representation of extreme behaviors. In case hypothesis **2-** is not fulfilled, such a sparse ‘profile’ can still be learned, but loses the low dimensional property of its supporting hyper-cubes. One major issue is that real data generally do not concentrate on sub-spaces of zero Lebesgue measure.

This is circumvented by setting to zero any coordinate less than a threshold $\epsilon > 0$, so that the corresponding ‘angle’ is assigned to a lower-dimensional face.

More formally, Figures 1.2 and 1.3 represent the transformed input space, resulting from classical standardization of the margins. After this non-linear transform, the representation of extreme data is linear and learned by estimating the mass on the sub-cones

$$\mathcal{C}_\alpha = \{\mathbf{v} \geq 0, \|\mathbf{v}\|_\infty \geq 1, v_j > 0 \text{ for } j \in \alpha, v_j = 0 \text{ for } j \notin \alpha\},$$

or more precisely, the mass of the angular measure Φ on the corresponding sub-spheres

$$\Omega_\alpha = \{\mathbf{x} \in S_\infty^{d-1} : x_i > 0 \text{ for } i \in \alpha, x_i = 0 \text{ for } i \notin \alpha\} = S_\infty^{d-1} \cap \mathcal{C}_\alpha,$$

represented in Figure 1.2.



FIGURE 1.2: Truncated cones in 3D

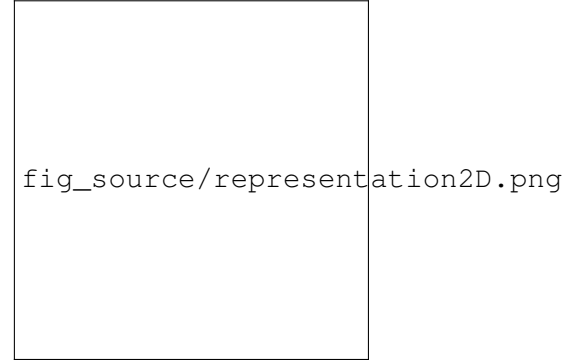


FIGURE 1.3: Truncated ϵ -cones in 2D

This is done using ϵ -thickened sub-cones $\mathcal{C}_\alpha^\epsilon$, corresponding to ϵ -thickened sub-spheres Ω_α^ϵ , as shown in Figure 1.3 in the two-dimensional case. We thus obtain an estimate $\widehat{\mathcal{M}}$ of the representation

$$\mathcal{M} = \{\Phi(\Omega_\alpha) : \emptyset \neq \alpha \subset \{1, \dots, d\}\}.$$

Theoretically, recovering the $(2^d - 1)$ -dimensional unknown vector \mathcal{M} amounts to roughly approximating the support of Φ using the partition $\{\Omega_\alpha, \alpha \subset \{1, \dots, d\}, \alpha \neq \emptyset\}$, that is, determine which Ω_α ’s have nonzero mass (and evaluating the mass $\Phi(\Omega_\alpha)$), or equivalently, which Φ_α ’s are nonzero. This support estimation is potentially sparse (if a small number of Ω_α have non-zero mass, *i.e.* Phenomenon **1**-) and potentially low-dimensional (if the dimensions of the sub-spheres Ω_α with non-zero mass are low, *i.e.* Phenomenon **2**-).

Anomaly Detection. Our proposed algorithm, DAMEX for Detecting Anomalies with Extremes, learns $\widehat{\mathcal{M}}$, a (possibly sparse and low-dimensional) representation of the angular measure, from which a scoring function can be defined in the context of anomaly detection. The underlying assumption is that an observation is potentially abnormal if its ‘direction’ (after a standardization of each marginal) is special regarding the other extreme observations. In other words, if it does not belong to the (sparse) representation $\widehat{\mathcal{M}}$. See Chapter 9 for details on how the scoring function is defined from this representation. According to the benchmarks derived

in this chapter, DAMEX significantly improves the performance (both in term of precision and of ROC curves) in extreme regions, inducing *e.g.* more vertical ROC curves near the origin.

Theoretical grounds. From the work on the STDF estimation summarized in the previous Section 1.6, in particular from Theorem 1.1 and from the ideas used to prove Theorem 1.2, we are able to derive some theoretical guaranties for this approach. Under non-restrictive assumptions standard in EVT (existence of the angular measure and continuous marginal c.d.f.), we obtain a non-asymptotic bound of the form

$$\sup_{\emptyset \neq \alpha \subset \{1, \dots, d\}} |\widehat{\mathcal{M}}(\alpha) - \mathcal{M}(\alpha)| \leq Cd \left(\sqrt{\frac{1}{\epsilon k} \log \frac{d}{\delta}} + Md\epsilon \right) + \text{bias}(\epsilon, k, n),$$

with probability greater than $1 - \delta$, where $k = k(n) \rightarrow \infty$ with $k(n) = o(n)$ can be interpreted as the number of data considered as extreme. The bias term goes to zero as $n \rightarrow \infty$, for any fixed ϵ .

1.8 Scikit-learn contributions

As an other contribution of this thesis, two classical anomaly detection algorithms, Isolation Forest and Local Outlier Factor have been implemented and merged on scikit-learn. These algorithms are presented in the Background Part, Section 2.2.

Scikit-learn, see Pedregosa et al. (2011), is an open-source library providing well-established machine learning methods. It is a Python module, the latter language being very popular for scientific computing, thanks to its high-level interactive nature. Scikit-learn provides a composition mechanism (through a *Pipeline* object) to combine estimators, preprocessing tools and model selection methods in such a way the user can easily construct complex ad-hoc algorithms. The development is done on *Github*¹, a Git repository hosting service which facilitates collaboration, as coding is done in strong interaction with other developers. Because of the large number of developers, emphasis is put on keeping the project maintainable, *e.g.* by avoiding duplicating code at the price of a reasonable loss of computational performance.

This contribution was supervised by Alexandre Gramfort and was funded by the Paris Saclay Center for Data Science. It also includes work for the scikit-learn maintenance like resolving issues and reviewing other contributors' pull requests.

1.9 Conclusion and Scientific Output

The contributions of this thesis can be summarized as follows.

¹<https://github.com/scikit-learn>

First, it proposes (Section 1.3.3) an adequate performance criterion, in order to compare possible candidate scoring function and to pick one eventually. The corresponding publication is Goix et al. (2015c):

- On Anomaly Ranking and Excess-Mass Curves. (AISTAT 2015).
Authors: Goix, Sabourin, and Cl  men  on.

However, the use of the Excess-Mass curve (just as Mass-Volume curve) to measure the quality of a scoring function s_n involves the computation of the Lebesgue measure $\text{Leb}(s_n \geq u)$, which is a major drawback for its use in high dimensional framework. Besides, these two criteria have originally been introduced to build scoring functions *via* Empirical Risk Minimization (ERM), and no study has been made on their use to evaluate scoring functions as ROC or PR criteria do.

As a second contribution, we show (empirically) that EM or MV based criteria are able to discriminate accurately (*w.r.t.* ROC or PR based criteria) between scoring functions in low dimension. Besides, we propose a methodology based on feature sub-sampling and aggregating to scale the use of EM or MV to higher dimensions. The corresponding publications are Goix (2016) and Goix & Thomas (2016):

- How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms? (ICML 2016, Workshop on Anomaly Detection).
Author: Goix.
- How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms? (submitted to NIPS 2016).
Authors: Goix and Thomas.

The third contribution of this thesis is to develop an efficient way of building accurate scoring functions. This is done by generalizing random forests to one-class classification. The corresponding publication is

- One-Class Splitting Criteria for Random Forests with Application to Anomaly Detection. (Submitted to NIPS 2016).
Authors: Goix, Brault, Drougard and Chiapino.

As a fourth contribution, we bring advances in multivariate EVT by providing non-asymptotic bounds for the estimation of the STDF, a functional characterizing the extreme dependence structure (Section 1.6). The corresponding publication is Goix et al. (2015b):

- Learning the dependence structure of rare events: a non-asymptotic study. (COLT 2015).
Authors: Goix, Sabourin, and Cl  men  on.

The fifth contribution is to design a statistical method that produces a (possibly sparse) representation of the dependence structure of extremes, while deriving non-asymptotic bounds to assess the accuracy of the estimation procedure (Section 1.7). This contribution also includes a multivariate EVT-based algorithm which returns a scoring functions defined in extremes region. This directly applies to anomaly detection as an abnormality score. The corresponding publications are Goix et al. (2016b), Goix et al. (2015a) and Goix et al. (2016c):

- Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. (AISTAT 2016 and NIPS 2015 Workshop on Nonparametric Methods for Large Scale Representation Learning).
Authors: Goix, Sabourin, and Cléménçon.
- Sparse Representation of Multivariate Extremes with Applications to Anomaly Detection. (Under review for Journal of Multivariate Analysis).
Authors: Goix, Sabourin, and Cléménçon.

As a last contribution (of incremental nature), two classical anomaly detection algorithms have been implemented and merged on scikit-learn. They are used in this dissertation for empirical comparison to attest the relevance of the forementioned approaches. The pull requests of these two contributions are available here:

- <https://github.com/scikit-learn/scikit-learn/pull/4163> (Isolation Forest)
- <https://github.com/scikit-learn/scikit-learn/pull/5279> (LOF)

Context of this work. This thesis was carried out in the STA (Statistiques et Applications) team of the Signal and Image Processing (TSI) department at Telecom ParisTech. The contributions presented in this thesis were supported by Ecole Normale Supérieure de Cachan via a ‘contrat doctoral pour normalien’ and by the industrial chair ‘Machine Learning for Big Data’ from Telecom ParisTech. The scikit-learn contributions have been supported by the Paris Saclay Center for Data Science regarding the collaboration with Alexandre Gramfort, and by the forementioned machine learning chair as regards the collaboration at New York University with Andreas Müller.

Outline of the thesis. This dissertation is organized as follows.

- Part I gathers the background work relevant to this thesis. Chapter 2 reviews classical anomaly detection algorithms through the scikit-learn library and presents implementative contributions of this thesis; Chapter 3 presents general results on measure concentration inequalities; and Chapter 4 provides a concise background on Extreme Value Theory.

- Part II deals with unsupervised anomaly detection criteria. Chapter 5 presents the details on anomaly ranking and excess-mass curve, as summarized above Section 1.3. Chapter 6 deals with the evaluation of anomaly detection algorithms, as summarized above Section 1.3. Chapter 1.4.
- Part III is about one-class random forests: Chapter 7 presents the details summarized above Section 1.5.
- Part IV focuses on EVT-based methods for anomaly detection. Chapter 8 deals with the stable tail dependence function as summarized above in Section 1.6. Chapter 9 describes how scoring functions can be build using EVT, as previously summarized in Section 1.7.

PART I

Preliminaries

Background on Anomaly Detection through Scikit-Learn

Chapter abstract In this chapter, we review some very classical anomaly detection algorithms and introduce the reader to the scikit-learn library. We also present relative implementative contributions of this thesis.

Note: The work on scikit-learn was supervised by Alexandre Gramfort and is the result of a collaboration with the Paris Saclay Center for Data Science. It includes the implementation of Isolation Forest (Section 2.2.4) and Local Outlier Factor (Section 2.2.3) algorithms, as well as a participation to the scikit-learn maintenance and pull requests review.

2.1 What is Anomaly Detection?

Anomaly Detection (and depending on the application domain, outlier detection, novelty detection, deviation detection, exception mining) generally consists in assuming that the dataset under study contains a *small* number of anomalies, generated by distribution models that *differ* from the one generating the vast majority of the data. This formulation motivates many statistical anomaly detection methods, based on the underlying assumption that anomalies occur in low probability regions of the data generating process. Here and hereafter, the term ‘normal data’ does not refer to Gaussian distributed data, but to *not abnormal* ones, *i.e.* data belonging to the above mentioned majority. Classical parametric techniques, like those developed in Barnett & Lewis (1994) or in Eskin (2000), assume that the normal data are generated by a distribution belonging to some specific, known in advance parametric model. The most popular non-parametric approaches include algorithms based on density (level set) estimation (see *e.g.* Schölkopf et al. (2001), Scott & Nowak (2006) or Breunig et al. (2000)), on dimensionality reduction (*cf* Shyu et al. (2003), Aggarwal & Yu (2001)) or on decision trees (Liu et al. (2008)). One may refer to Hodge & Austin (2004), Chandola et al. (2009), Patcha & Park (2007) and Markou & Singh (2003) for excellent overviews of current research on Anomaly Detection, ad-hoc techniques being far too numerous to be listed here in an exhaustive manner.

Most usual anomaly detection algorithms actually provide more than a predicted label for any new observation, abnormal vs. normal. Instead, they return a real valued function, termed a *scoring function*, defining a preorder/ranking on the input space. Such a function permits to

rank any observations according to their supposed ‘degree of abnormality’ and thresholding it yields a decision rule that splits the input space into ‘normal’ and ‘abnormal’ regions. In various fields (*e.g.* fleet management, monitoring of energy/transportation networks), when confronted with massive data, being able to rank observations according to their degree of abnormality may significantly improve operational processes and allow for a prioritization of actions to be taken, especially in situations where human expertise is required to check each observation is time-consuming.

From a machine learning perspective, anomaly detection can be considered as a specific classification task, where the usual assumption in supervised learning stipulating that the dataset contains structural information regarding all classes breaks down, see Roberts (1999). This typically happens in the case of two highly unbalanced classes: the normal class is expected to regroup a large majority of the dataset, so that the very small number of points representing the abnormal class does not allow to learn information about this class. In a clustering based approach, it can be interpreted as the presence of a single cluster, corresponding to the normal data. The abnormal ones are too limited to share a common structure, *i.e.* to form a second cluster. Their only characteristic is precisely to lie outside the normal cluster, namely to lack any structure. Thus, common classification approaches may not be applied as such, even in a supervised context. **Supervised** anomaly detection consists in training the algorithm on a labeled (normal/abnormal) dataset including both normal and abnormal observations. In the **semi-supervised** context, only normal data are available for training. This is the case in applications where normal operations are known but intrusion/attacks/viruses are unknown and should be detected. In the **unsupervised** setup, no assumption is made on the data which consist in unlabeled normal and abnormal instances. In general, a method from the semi-supervised framework may apply to the unsupervised one, as soon as the number of anomalies is sufficiently weak to prevent the algorithm from fitting them when learning the normal behavior. Such a method should be robust to outlying observations.

As mentioned in the introduction, contribution of this thesis includes the implementation of two classical anomaly detection algorithms on the open-source scikit-learn library (Pedregosa et al. (2011)), namely the Isolation Forest algorithm (Liu et al. (2008)) and the Local Outlier Factor algorithm (Breunig et al. (2000)), which are respectively presented in sections 2.2.3 and 2.2.4. This work was supervised by Alexandre Gramfort and is the result of a collaboration with the Paris Saclay Center for Data Science. It also includes participation to the scikit-learn maintenance and pull requests review.

The following section provides insights on Anomaly Detection through scikit-learn by describing and comparing anomaly detection algorithms from this library. Part of this section are modified versions of the documentation included in the forementioned scikit-learn contribution.

2.2 Three efficient Anomaly Detection Algorithms

2.2.1 What is Scikit-learn?

Scikit-learn, see Pedregosa et al. (2011), is an open-source library which provides well-established machine learning methods. It is a Python module, the latter language being very popular for scientific computing, thanks to its high-level interactive nature. Python is enjoying this recent years a strong expansion both in academic and industrial settings. Scikit-learn takes advantage of this favourable backdrop and extends this general-purpose programming language with machine learning operation: It not only provides implementations of many established algorithms, both supervised and unsupervised, while keeping an easy-to-use interface tightly integrated with the Python language. But it also provides a composition mechanism (through a *Pipeline* object) to combine estimators, preprocessing tools and model selection methods in such a way the user can easily constructs complex ad-hoc algorithms.

Scikit-learn depends only on *numpy* (the base data structure used for data and model parameters, see Van Der Walt et al. (2011)) and *scipy* (to handle common numerical operations, see Jones et al. (2015)). Most of the Scikit-learn package is written in python and *cython*, a compiled programming language for combining C in Python to achieve the performance of C with high-level programming in Python-like syntax.

The development is done on *github*¹, a Git repository hosting service which facilitates collaboration, as coding is done in strong interaction with other developpers. Because of the large number them, emphasis is put on keeping the project maintainable, *e.g.* by avoiding duplicating code.

Scikit-learn benefits from a simple and consistent API (Application Programming Interface), see Buitinck et al. (2013), through the *estimator* interface. This interface is followed by all (supervised and unsupervised) learning algorithms as well as other tasks such as preprocessing, feature extraction and selection. The central object *estimator* implements a *fit* method to learn from training data, taking as argument an input data array (and optionnally an array of labels for supervised problems). The initialization of the estimator is done separately, before training, in such a way the constructor doesn't see any data and can be seen as a function taking as input the model hyper-parameters and returning the learning algorithm initialized with these parameters. Relevant default parameters are provided for each algorithm. To illustrate initialization and fit steps, the snippet below considers an anomaly detection learning task with the *Isolation Forest* algorithm.

¹<https://github.com/scikit-learn>

```
# Import the IsolationForest algorithm from the ensemble module
from sklearn.ensemble import IsolationForest

# Instantiate with specified hyper-parameters
IF = IsolationForest(n_trees=100, max_samples=256)

# Fit the model on training data (build the trees of the forest)
IF.fit(X_train)
```

In this code example, the Isolation Forest algorithm is imported from the *ensemble* module of scikit-learn, which contains the ensemble-based estimators such as bagging or boosting methods. Then, an *IsolationForest* instance *IF* is initialized with a number of trees of 100 (see Section 2.2.4 for details on this algorithm). Finally, the model is learned from training data *X_train* and stored on the *IF* object for later use. Since all estimators share the same API, it is possible to train a Local Outlier Factor algorithm by simply replacing the constructor name *IsolationForest*(*n_trees* = 100) in the snippet above by *LocalOutlierFactor*().

Some estimators (such as supervised estimators or some of the unsupervised ones, like Isolation Forest and LOF algorithm) are called *predictors* and implement a *predict* method that takes a data array and returns predictions (labels or values computed by the model). Other estimators (e.g. PCA) are called *transformer* and implement a *transform* method returning modified input data. The following code example illustrates how simple it is to predict labels with the predictor interface. It suffices to add the line of code below to the previous snippet.

```
# Perform prediction on new data
y_pred = IF.predict(X_test)
# Here y_pred is a vector of binary labels (+1 if inlier, -1 if abnormal)
```

2.2.2 One-class SVM

The SVM algorithm is essentially a two-class algorithm (*i.e.* one needs negative as well as positive examples). Schölkopf et al. (2001) extended the SVM methodology to handle training using only positive information: the One-Class Support Vector Machine (OCSVM) treats the origin as the only member of the second class (after mapping the data to some feature space). Thus the OCSVM finds a separating hyperplane between the origin and the mapped one class.

The OCSVM consists in estimating Minimum Volume sets, which amounts (if the density has no flat parts) to estimating density level sets, as mentioned in the introduction. In Vert & Vert (2006), it is shown that the OCSVM is a consistent estimator of density level sets, and that the solution function returned by the OCSVM gives an estimate of the tail of the underlying density.

Figures 2.1 and 2.2 summarizes the theoretical insights of OCSVM compared to the standard SVM, respectively for the hard-margin (no error is tolerated during training) and soft-margin separation (some margin errors are tolerated in training).

SVM	OCSVM
$\min_{w,b} \frac{1}{2} \ w\ ^2$ $\text{s.t. } \forall i, y_i(\langle w, x_i \rangle + b) \geq 1$	$\min_w \frac{1}{2} \ w\ ^2$ $\text{s.t. } \forall i, \langle w, x_i \rangle \geq 1$
<div>fig_source/OCSVM_hard.png</div>	
<p>decision function:</p> $f(x) = \text{sgn}(\langle w, x \rangle + b) \text{ (red line)}$	<p>decision function:</p> $f(x) = \text{sgn}(\langle w, x \rangle - 1) \text{ (green line)}$
<p>-Lagrange multipliers: α_i ($\alpha_i > 0$ when the constraint is an equality for x_i)</p> <p>-Support vectors: $\text{SV} = \{x_i, \alpha_i > 0\}$</p> <p>-Margin errors: $\text{ME} = \emptyset$</p>	
$w = \sum_i \alpha_i y_i x_i$	$w = \sum_i \alpha_i x_i$

TABLE 2.1: SVM vs. OCSVM (hard-margin separation)

In the ν -soft margin separation framework, letting Φ be the mapping function determined by a kernel function k (i.e. $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$), the separating hyperplane defined w.r.t. a vector w and an offset ρ is given by the solution of

$$\begin{aligned} \min_{w, \xi, \rho} \quad & \frac{1}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^n \xi_i - \nu \rho \\ \text{s.t.} \quad & \langle w, \Phi(x_i) \rangle \geq \rho - \xi_i, \quad 1 \leq i \leq n \\ & \xi_i \geq 0, \end{aligned}$$

SVM	OCSVM
$\min_{w, \xi, \rho, b} \frac{1}{2} \ w\ ^2 + \frac{1}{n} \sum_{i=1}^n \xi_i - \nu \rho$ $\text{s.t. } \forall i, y_i(\langle w, x_i \rangle + b) \geq \rho - \xi_i$ $\xi_i \geq 0$	$\min_{w, \xi, \rho} \frac{1}{2} \ w\ ^2 + \frac{1}{n} \sum_{i=1}^n \xi_i - \nu \rho$ $\text{s.t. } \forall i, \langle w, x_i \rangle \geq \rho - \xi_i$ $\xi_i \geq 0$
<div>fig_source/OCSVM_soft.png</div>	
<p>decision function:</p> $f(x) = \text{sgn}(\langle w, x \rangle + b) \text{ (red line)}$	<p>decision function:</p> $f(x) = \text{sgn}(\langle w, x \rangle - \rho) \text{ (green line)}$
<p>-Lagrange multipliers: α_i, β_i (one for each constraint, $\beta_i > 0$ when $\xi_i = 0$)</p> <p>-Support vectors: $\text{SV} = \{x_i, \alpha_i > 0\}$</p> <p>-Margin errors: $\text{ME} = \{x_i, \xi_i > 0\} = \{x_i, \beta_i > 0\}$ (for OCSVM, ME=anomalies)</p> <p>-$\text{SV} \setminus \text{ME} = \{x_i, \alpha_i, \beta_i > 0\}$</p>	
$w = \sum_i \alpha_i y_i x_i$	$w = \sum_i \alpha_i x_i$
$\frac{ \text{ME} }{n} \leq \nu \leq \frac{ \text{SV} }{n}$ $\rho = \langle w, x_i \rangle \quad \forall x_i \in \text{SV} \setminus \text{ME}$	

TABLE 2.2: SVM vs. OCSVM (ν -soft margin separation)

where ν is previously set. An interesting fact is that ν is an upper bound on the fraction of outliers and a lower bound on the fraction of support vectors, both of which converging to ν almost surely as $n \rightarrow \infty$ (under some continuity assumption). Then, the empirical mass of the estimated level set is greater than $1 - \nu$ and converges almost surely to $1 - \nu$ as n tends to infinity. Hence one usual approach is to choose $\nu = 1 - \alpha$ to estimate a MV-set with mass (at least) α . For insights on the calibration of One-Class SVM, see for instance Thomas et al. (2015). The OCSVM is mainly applied with Gaussian kernels and its performance highly depends on the kernel bandwidth selection. The complexity of OCSVM training is the same as for the standard SVM, namely $O(n^3d)$ where n is the number of samples and d the dimension of the input space. However, one can often expect a complexity of $O(n^2d)$, see Bottou & Lin (2007). From its linear complexity *w.r.t.* the number of features d , OCSVM scales well in large dimension, and performance remains good even when the dimension is greater than n . By using only a small subset of the training dataset (support vectors) in the decision function, it is memory efficient. However, OCSVM suffers from practical limitation: 1) the non-linear training complexity in the number of observations, which limits its use on very large datasets; 2) its sensitivity to the parameter ν and to the kernel bandwidth, which makes calibration tricky; 3) parametrization of the mass of the MV set estimated by the OCSVM via the parameter ν does not allow to obtain nested set estimates as the mass α increases.

2.2.3 Local Outlier Factor algorithm

One other very efficient way of performing outlier detection in datasets whose dimension is moderately large is to use the Local Outlier Factor (LOF) algorithm proposed in Breunig et al. (2000).

This algorithm computes a score reflecting the degree of abnormality of the observations, the so-called local outlier factor. It measures the local deviation of a given data point with respect to its neighbors. By comparing the local density near a sample to the local densities of its neighbors, one can identify points which have a substantially lower density than their neighbors. These are considered to be outliers.

In practice the local density is obtained from the k -nearest neighbors. The LOF score of an observation is equal to the ratio of the average local density of his k -nearest neighbors, and his own local density: a normal instance is expected to have a local density similar to that of its neighbors, while abnormal data are expected to have much smaller local density.

The strength of the LOF algorithm is that it takes both local and global properties of datasets into consideration: it can perform well even in datasets where abnormal samples have different underlying densities. The question is not, how isolated the sample is, but how isolated it is with respect to the surrounding neighborhood.

This strategy is illustrated in the code example below, returning Figure 2.1.



fig_source/lof.png

FIGURE 2.1: LOF example

```
"""
=====
Anomaly detection with Local Outlier Factor (LOF)
=====

This example uses the LocalOutlierFactor estimator
for anomaly detection.
"""

import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import LocalOutlierFactor

np.random.seed(42)

# Generate train data
X = 0.3 * np.random.randn(100, 2)
X_train = np.r_[X + 2, X - 2]
# Generate some regular novel observations
X = 0.3 * np.random.randn(20, 2)
X_test = np.r_[X + 2, X - 2]
# Generate some abnormal novel observations
X_outliers = np.random.uniform(low=-4, high=4, size=(20, 2))

# fit the model
clf = LocalOutlierFactor()
clf.fit(X_train)
y_pred_train = clf.predict(X_train)
y_pred_test = clf.predict(X_test)
y_pred_outliers = clf.predict(X_outliers)
```

```

# plot the line, the samples, and the nearest vectors to the plane
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)

plt.title("Local Outlier Factor (LOF)")
plt.contourf(xx, yy, Z, cmap=plt.cm.Blues_r)

b1 = plt.scatter(X_train[:, 0], X_train[:, 1], c='white')
b2 = plt.scatter(X_test[:, 0], X_test[:, 1], c='green')
c = plt.scatter(X_outliers[:, 0], X_outliers[:, 1], c='red')
plt.axis('tight')
plt.xlim((-5, 5))
plt.ylim((-5, 5))
plt.legend([b1, b2, c],
           ["training observations",
            "new regular observations", "new abnormal observations"],
           loc="upper left")
plt.show()

```

2.2.4 Isolation Forest

One efficient way of performing outlier detection in high-dimensional datasets is to use random forests. The IsolationForest proposed in Liu et al. (2008) 'isolates' observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node. This path length, averaged over a forest of such random trees, is a measure of abnormality and our decision function. Random partitioning produces noticeable shorter paths for anomalies, see figures 2.2 and 2.3. Moreover, the average depth of a sample over the forest seems to converge to some limits, the latter being different depending on if the sample is or not an anomaly. Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies. This strategy is illustrated in the code example below returning Figure 2.4.

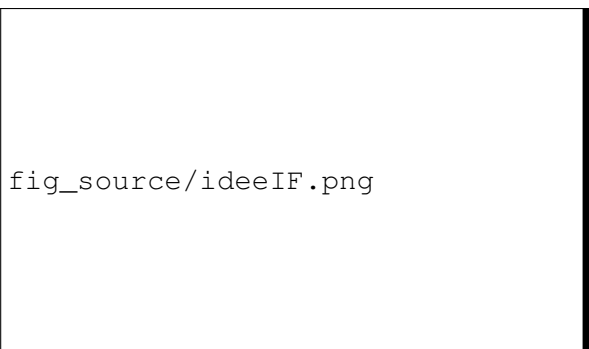


FIGURE 2.2: Anomalies are isolated more quickly

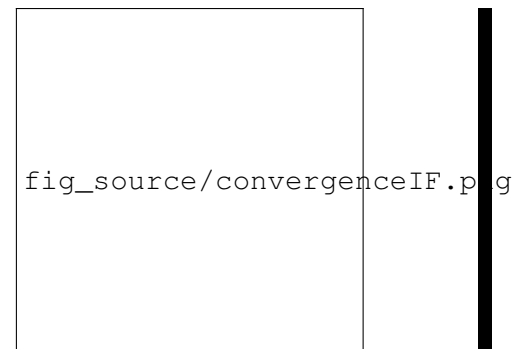


FIGURE 2.3: Convergence of the averaged depth



FIGURE 2.4: Isolation Forest example

```

"""
=====
IsolationForest example
=====

An example using IsolationForest for anomaly detection.

"""

import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest

rng = np.random.RandomState(42)

# Generate train data
X = 0.3 * rng.randn(100, 2)
X_train = np.r_[X + 2, X - 2]
# Generate some regular novel observations
X = 0.3 * rng.randn(20, 2)
X_test = np.r_[X + 2, X - 2]
# Generate some abnormal novel observations
X_outliers = rng.uniform(low=-4, high=4, size=(20, 2))

# fit the model
clf = IsolationForest(max_samples=100, random_state=rng)
clf.fit(X_train)
y_pred_train = clf.predict(X_train)
y_pred_test = clf.predict(X_test)
y_pred_outliers = clf.predict(X_outliers)

```



```

# plot the line, the samples, and the nearest vectors to the plane
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)

plt.title("IsolationForest")
plt.contourf(xx, yy, Z, cmap=plt.cm.Blues_r)

b1 = plt.scatter(X_train[:, 0], X_train[:, 1], c='white')
b2 = plt.scatter(X_test[:, 0], X_test[:, 1], c='green')
c = plt.scatter(X_outliers[:, 0], X_outliers[:, 1], c='red')
plt.axis('tight')
plt.xlim((-5, 5))
plt.ylim((-5, 5))
plt.legend([b1, b2, c],
           ["training observations",
            "new regular observations", "new abnormal observations"],
           loc="upper left")
plt.show()

```

As a conclusion of this Section, Figures 2.5, 2.6 and 2.7 draw a comparison of the three anomaly detection algorithms introduced in this section:

- the One-Class SVM is able to capture the shape of the data set, hence performing well when the data is strongly non-Gaussian, i.e. with two well-separated clusters;
- the Isolation Forest algorithm, is adapted to large-dimensional settings, even if it performs quite well in the examples below.
- the Local Outlier Factor measures the local deviation of a given data point with respect to its neighbors by comparing their local density.

The ground truth about inliers and outliers is given by the points colors while the orange-filled area indicates which points are reported as inliers by each method.

Here, we assume that we know the fraction of outliers in the datasets. Thus rather than using the ‘predict’ method of the objects, we set the threshold on the decision function to separate out the corresponding fraction. Anomalies are uniformly drawn according to an uniform distribution.



fig_source/ADcomparison1.png

FIGURE 2.5: gaussian normal data with one single mode



fig_source/ADcomparison2.png

FIGURE 2.6: gaussian normal data with two modes



fig_source/ADcomparison3.png

FIGURE 2.7: gaussian normal data with two strongly separate modes

```

"""
=====
Outlier detection with several methods.
=====
"""

import numpy as np
import matplotlib.pyplot as plt
import matplotlib.font_manager
from scipy import stats

from sklearn import svm
from sklearn.covariance import EllipticEnvelope
from sklearn.ensemble import IsolationForest
from sklearn.neighbors import LocalOutlierFactor

rng = np.random.RandomState(42)

# Example settings
n_samples = 200
outliers_fraction = 0.25
clusters_separation = [0, 1, 2]

# define two outlier detection tools to be compared
classifiers = {
    "One-Class SVM": svm.OneClassSVM(nu=0.95 * outliers_fraction + 0.05,
                                     kernel="rbf", gamma=0.1),
    "robust covariance estimator": EllipticEnvelope(contamination=.25),
    "Isolation Forest": IsolationForest(random_state=rng),
    "Local Outlier Factor": LocalOutlierFactor(n_neighbors=35, contamination=0.25)}

# Compare given classifiers under given settings
xx, yy = np.meshgrid(np.linspace(-7, 7, 100), np.linspace(-7, 7, 100))
n_inliers = int((1. - outliers_fraction) * n_samples)
n_outliers = int(outliers_fraction * n_samples)
ground_truth = np.ones(n_samples, dtype=int)
ground_truth[-n_outliers:] = 0

# Fit the problem with varying cluster separation
for i, offset in enumerate(clusters_separation):
    np.random.seed(42)
    # Data generation
    X1 = 0.3 * np.random.randn(n_inliers // 2, 2) - offset
    X2 = 0.3 * np.random.randn(n_inliers // 2, 2) + offset
    X = np.r_[X1, X2]
    # Add outliers
    X = np.r_[X, np.random.uniform(low=-6, high=6, size=(n_outliers, 2))]

    # Fit the model
    plt.figure(figsize=(10, 5))
    for i, (clf_name, clf) in enumerate(classifiers.items()):
        # fit the data and tag outliers
        clf.fit(X)
        y_pred = clf.decision_function(X).ravel()
        threshold = stats.scoreatpercentile(y_pred,
                                           100 * outliers_fraction)

        y_pred = y_pred > threshold
        n_errors = (y_pred != ground_truth).sum()

```

```

# plot the levels lines and the points
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
subplot = plt.subplot(1, 3, i + 1)
subplot.contourf(xx, yy, Z, levels=np.linspace(Z.min(), threshold, 7),
                 cmap=plt.cm.Blues_r)
a = subplot.contour(xx, yy, Z, levels=[threshold],
                   linewidths=2, colors='red')
subplot.contourf(xx, yy, Z, levels=[threshold, Z.max()],
                 colors='orange')
b = subplot.scatter(X[:-n_outliers, 0], X[:-n_outliers, 1], c='white')
c = subplot.scatter(X[-n_outliers:, 0], X[-n_outliers:, 1], c='black')
subplot.axis('tight')
subplot.legend(
    [a.collections[0], b, c],
    ['learned decision function', 'true inliers', 'true outliers'],
    prop=matplotlib.font_manager.FontProperties(size=10),
    loc='lower right')
subplot.set_xlabel("%d. %s (errors: %d)" % (i + 1, clf_name, n_errors))
subplot.set_xlim((-7, 7))
subplot.set_ylim((-7, 7))
plt.subplots_adjust(0.04, 0.1, 0.96, 0.94, 0.1, 0.26)
plt.suptitle("Outlier detection")

plt.show()

```

CHAPTER 3

Concentration Inequalities from the Method of bounded differences

Chapter abstract This chapter presents general results on measure concentration inequalities, obtained via martingal methods or with Vapnik-Chervonenkis theory. In the last section 3.1.3 of this chapter, a link is also made with contributions presented in Chapter 8 which builds on some concentration inequality stated and proved here.

We recommend McDiarmid (1998) and Janson (2002) for good references on this subject, and Boucheron et al. (2013) for a complete review on concentration inequalities.

3.1 Two fundamental results

The two theorems 3.4 and 3.7 derived in this section are very powerful and allow to derive lots of classical concentration inequalities, like Hoeffding, Azuma, Bernstein or McDiarmid inequalities. The first one applies to bounded *r.v.* while the second one only makes variance assumption.

3.1.1 Preliminary definitions

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let X be a random variable on this space and \mathcal{G} a sub- σ -algebra of \mathcal{F} .

Definition 3.1. Assume X is a real *r.v.*. We define $\sup(X|\mathcal{G})$ as the unique real *r.v.* $f : \Omega \rightarrow \mathbb{R}$ verifying:

- (i) f is \mathcal{G} -measurable
- (ii) $X \leq f$ a.s.
- (iii) If $g : \Omega \rightarrow \mathbb{R}$ verifies (i) and (ii) then $f \leq g$ a.s.

Note that we clearly have $\sup(X|\mathcal{G}) \geq \mathbb{E}(X|\mathcal{G})$ and $\sup(X|\mathcal{G}_1) \geq \sup(X|\mathcal{G}_2)$ when $\mathcal{G}_1 \subset \mathcal{G}_2$.

Definition 3.2. Assume X is bounded. Let $(\mathcal{F}_k)_{0 \leq k \leq n}$ be a filtration of \mathcal{F} such that X is \mathcal{F}_n -mesurable. We denote X_1, \dots, X_n the martingale $X_k = \mathbb{E}(X|\mathcal{F}_k)$ and $Y_k = X_k - X_{k-1}$ the associated martingale difference. The r.v. $\mathbf{ran}(X|\mathcal{G}) := \sup(X|\mathcal{G}) + \sup(-X|\mathcal{G})$ is called the conditional range of X w.r.t. \mathcal{G} . Then we define:

- * $\mathbf{ran}_k = \mathbf{ran}(Y_k|\mathcal{F}_{k-1}) = \mathbf{ran}(X_k|\mathcal{F}_{k-1})$ the conditional range,
- * $\mathbf{R}^2 = \sum_1^n \mathbf{ran}_k^2$ the sum of squared conditional ranges, and $\hat{\mathbf{r}}^2 = \supess(\mathbf{R}^2)$ the maximum sum of squared conditional ranges.

Definition 3.3. We place ourselves in the same context as in the previous definition, but without assuming X is bounded. The r.v. $\mathbf{var}(X|\mathcal{G}) := \mathbb{E}((X - \mathbb{E}(X|\mathcal{G}))^2|\mathcal{G})$ is called the conditional variance of X w.r.t. \mathcal{G} . Then we define:

- $\mathbf{var}_k = \mathbf{var}(Y_k|\mathcal{F}_{k-1}) = \mathbf{var}(X_k|\mathcal{F}_{k-1})$ the conditional variance,
- $\mathbf{V} = \sum_1^n \mathbf{var}_k$ the sum of conditional variances and $\hat{\mathbf{v}} = \supess(\mathbf{V})$ the maximum sum of conditional variances,
- * $\mathbf{dev}_k^+ = \sup(Y_k|\mathcal{F}_{k-1})$ the conditional positive deviation,
- * $\mathbf{maxdev}^+ = \supess(\max_{0 \leq k \leq n} \mathbf{dev}_k^+)$ the maximum conditional positive deviation.

The r.v. V is also called the ‘predictable quadratic variation’ of the martingale (X_k) and is such that $\mathbb{E}(V) = \mathbf{var}(X)$.

3.1.2 Inequality for Bounded Random Variables

Theorem 3.4. Let X a bounded r.v. with $\mathbb{E}(X) = \mu$, and $(\mathcal{F}_k)_{0 \leq k \leq n}$ a filtration of \mathcal{F} such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and such that X is \mathcal{F}_n -mesurable. Then for any $t \geq 0$,

$$\mathbb{P}(X - \mu \geq t) \leq e^{-2t^2/\hat{\mathbf{r}}^2},$$

and more generally

$$\forall r^2 \geq 0, \quad \mathbb{P}((X - \mu \geq t) \cap (\mathbf{R}^2 \leq r^2)) \leq e^{-2t^2/r^2}.$$

To prove this result we need the two following lemmas.

Lemma 3.5. Let $(\mathcal{F}_k)_{0 \leq k \leq n}$ be a filtration of \mathcal{F} with $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and $(Y_k)_{1 \leq k \leq n}$ be a martingale difference for this filtration such that each Y_k is bounded. Let Z be any random variable. Then:

$$\mathbb{E}(Ze^{h \sum Y_k}) \leq \sup(Z \prod_{k=1}^n \mathbb{E}(e^{h Y_k}|\mathcal{F}_{k-1}))$$

Proof. This result can be easily proved by induction.

$$\begin{aligned}
\mathbb{E} \left[Z e^{h \sum Y_k} \right] &= \mathbb{E} \left[e^{h Y_1} \mathbb{E} \left[Z e^{h \sum_{2}^n Y_k} \mid \mathcal{F}_1 \right] \right] \\
&= \mathbb{E} \left[e^{h Y_1} \mathbb{E} \left[e^{h Y_2} \dots \mathbb{E} \left[\mathbb{E} [Z \mid \mathcal{F}_n] e^{h Y_n} \mid \mathcal{F}_{n-1} \right] \dots \mid \mathcal{F}_1 \right] \right] \\
&\leq \mathbb{E} \left[e^{h Y_1} \mathbb{E} \left[e^{h Y_2} \dots \mathbb{E} \left[\sup [Z \mid \mathcal{F}_n] e^{h Y_n} \mid \mathcal{F}_{n-1} \right] \dots \mid \mathcal{F}_1 \right] \right] \\
&= \mathbb{E} \left[e^{h Y_1} \mathbb{E} \left[e^{h Y_2} \dots \sup \left[Z \mathbb{E} \left[e^{h Y_n} \mid \mathcal{F}_{n-1} \right] \mid \mathcal{F}_n \right] \dots \mid \mathcal{F}_1 \right] \right] \\
&= \sup \left[Z \prod_k \mathbb{E}(e^{h Y_k} \mid \mathcal{F}_{k-1}) \mid \mathcal{F}_n \right] \\
&\leq \sup \left[Z \prod_k \mathbb{E}(e^{h Y_k} \mid \mathcal{F}_{k-1}) \right] \quad (\text{since } \mathcal{F}_0 \subset \mathcal{F}_n)
\end{aligned}$$

□

This lemma can be interpreted as doing ‘almost as if’ $\sum Y_k$ was a sum of independent variables.

Lemma 3.6. *Let X be a random variable such that $\mathbb{E}(X) = 0$ and $a \leq X \leq b$, then for any $h > 0$, we have $\mathbb{E}(e^{hX}) \leq e^{\frac{1}{8}h^2(b-a)^2}$. This result remains true with conditional expectation.*

Proof. The proof of this result does not present any difficulty but is quite technical. It is based on the convexity of the function $x \mapsto e^{hx}$ (see McDiarmid (1998) for details). □

Proof of Theorem 3.4. This proof follows a traditional scheme, based on four steps: exponential Markov inequality introducing a parameter h ; decomposition of the exponential term using independence (or in the present case using Lemma 3.5 which plays the same role); upper bound on each term with Lemma 3.6; and finally optimization in parameter h .

Let $X_k = \mathbb{E}(X \mid \mathcal{F}_{k-1})$ and $Y_k = X_k - X_{k-1}$ the associated martingale difference. Define the r.v. Z as $Z = \mathbb{1}_{R^2 \leq r^2}$. Exponential Markov inequality yields, for any $h > 0$,

$$\begin{aligned}
\mathbb{P}((X - \mu \geq t) \cap (R^2 \leq r^2)) &= \mathbb{P}(Z e^{h(X-\mu)} \geq e^{ht}) \\
&\leq e^{-ht} \mathbb{E}(Z e^{h(X-\mu)}) \\
&\leq e^{-ht} \mathbb{E}(Z e^{h(\sum Y_k)})
\end{aligned}$$

From Lemma 3.6, $\mathbb{E}(e^{h Y_k} \mid \mathcal{F}_{k-1}) \leq e^{\frac{1}{8}h^2 r_k^2}$ so that using Lemma 3.5,

$$\begin{aligned}
\mathbb{E}(Z e^{h \sum Y_k}) &\leq \sup(Z \prod \mathbb{E}(e^{h Y_k} \mid \mathcal{F}_{k-1})), \\
&\leq \sup(Z \prod e^{\frac{1}{8}h^2 r_k^2}), \\
&= \sup(Z e^{\frac{1}{8}h^2 R^2}), \\
&\leq e^{\frac{1}{8} \sup(Z R^2)}, \\
&\leq e^{\frac{1}{8}h^2 r^2}
\end{aligned}$$

By setting $h = \frac{4t}{r^2}$, we finally obtain

$$\mathbb{P}((X - \mu \geq t) \cap (R^2 \leq r^2)) \leq e^{-ht + \frac{1}{8}h^2r^2} \leq e^{-2t^2/r^2}.$$

□

3.1.3 Bernstein-type Inequality (with variance term)

Theorem 3.7. *Let X be a r.v. with $\mathbb{E}(X) = \mu$ and $(\mathcal{F}_k)_{0 \leq k \leq n}$ a filtration of \mathcal{F} such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and such that X is \mathcal{F}_n -measurable. Let $b = \max dev^+$ the maximum conditional deviation assumed to be finite, and $\hat{v} = \sup V$ the maximum sum of conditional variances also assumed to be finite. Then, for any $t \geq 0$,*

$$\mathbb{P}(X - \mu \geq t) \leq e^{-\frac{t^2}{2(\hat{v} + bt/3)}},$$

and more generally, for any $v \geq 0$,

$$\mathbb{P}((X - \mu \geq t) \cap (V \leq v)) \leq e^{-\frac{t^2}{2(v + bt/3)}}.$$

Unlike Theorem 3.4, this result also applies in the case of unbounded r.v. X . Note that even in the case X is bounded, Theorem 3.7 may give better bounds if the variance term \hat{v} is small enough.

To prove this result, two lemmas are needed: Lemma 3.5 previously stated, exploiting the decomposition into martingale differences and thus playing the same role as independence; and the following lemma replacing Lemma 3.6 in the case of non-necessarily bounded r.v., but with bounded variance.

Lemma 3.8. *Let g be the non-increasing functional defined for $x \neq 0$ by $g(x) = \frac{e^x - 1 - x}{x^2}$, and X a r.v. satisfying $\mathbb{E}(X) = 0$ and $X \leq b$. Then $\mathbb{E}(e^X) \leq e^{g(b)var(X)}$, and this result still holds with conditional expectation and variance, and replacing b by the associated conditional supremum.*

Proof. Noting that $e^x \leq 1 + x + x^2g(b)$ for $x \leq b$, we have $\mathbb{E}(e^X) \leq 1 + g(b)var(X) \leq e^{g(b)var(X)}$. □

Proof of Theorem 3.7. The proof follows the same classical lines as the one of Theorem 3.4. Let Y_1, \dots, Y_n be the martingale differences associated to X and (\mathcal{F}_k) , and $Z = \mathbb{1}_{V \leq v}$. Exponential

Markov inequality yields, for every $h > 0$,

$$\begin{aligned}\mathbb{P}((X - \mu \geq t) \cap (V \leq v)) &= \mathbb{P}(Ze^{h(X-\mu)} \geq e^{ht}) \\ &\leq e^{-ht} \mathbb{E}(Ze^{h(X-\mu)}) \\ &\leq e^{-ht} \mathbb{E}(Ze^{h(\sum Y_k)})\end{aligned}$$

From Lemma 3.8, $\mathbb{E}(e^{hY_k} | \mathcal{F}_{k-1}) \leq e^{h^2 g(h \text{dev}_k^+) \text{var}_k} \leq e^{h^2 g(hb) \text{var}_k}$ so that from Lemma 3.5 we obtain,

$$\begin{aligned}\mathbb{E}(Ze^{h \sum Y_k}) &\leq \sup(Z \prod \mathbb{E}(e^{hY_k} | \mathcal{F}_{k-1})) \\ &\leq \sup(Z \prod e^{h^2 g(hb) \text{var}_k}) \\ &= \sup(Z e^{h^2 g(hb) V}) \\ &\leq e^{h^2 g(hb) \sup(ZV)} \\ &\leq e^{h^2 g(hb) v}.\end{aligned}$$

By setting $h = \frac{1}{b} \ln(1 + \frac{bt}{v})$ and using the fact that for every positive x , we have $(1+x) \ln(1+x) - x \geq 3x^2/(6+2x)$, we finally get

$$\begin{aligned}\mathbb{P}((X - \mu \geq t) \cap (R^2 \leq r^2)) &\leq e^{-ht + h^2 g(hb) v} \\ &\leq e^{-\frac{t^2}{2(v+bt/3)}}.\end{aligned}$$

□

3.2 Famous Inequalities

In this section, we illustrate the strength of Theorem 3.4 and Theorem 3.7 by deriving as corollaries classical concentration inequalities. The first three propositions hold for bounded random variable and derive from Theorem 3.4. The last one (Bernstein) holds under variance assumption and derive from Theorem 3.7.

Proposition 3.9. (AZUMA-HOEFFDING INEQUALITY) *Let $(\mathcal{F}_k)_{0 \leq k \leq n}$ be a filtration of \mathcal{F} such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$, Z a martingale and Y the associated martingale difference. If for every k , $|Y_k| \leq c_k$, then we have*

$$\mathbb{P}(\sum_{k=1}^n Y_k \geq t) \leq e^{-\frac{t^2}{2 \sum_{k=1}^n c_k^2}}.$$

Moreover, the same inequality holds when replacing $\sum_{k=1}^n Y_k$ by $-\sum_{k=1}^n Y_k$.

Proof. Apply Theorem 3.4 with $X = \sum_1^n Y_k$, $\mathcal{F}_k = \sigma(Y_1, \dots, Y_k)$ and $X_k = \mathbb{E}(X | \mathcal{F}_k)$. Thus, $\mu = 0$, $X_k = \sum_1^k Y_i$ because Z is a martingale, and $Y_i = X_i - X_{i-1}$. Therefore, $\text{ran}_k =$

$\text{ran}(Y_k|\mathcal{F}_k) \leq 2c_k$, hence $R^2 \leq 4 \sum c_k^2$ and $\hat{r}^2 \leq 4 \sum c_k^2$. By Theorem 3.4, $\mathbb{P}(X \geq t) \leq e^{\frac{-2t^2}{\hat{r}^2}} \leq e^{-\frac{t^2}{2 \sum c_k^2}}$. Applying this inequality to $-X$, we obtain the desired result. \square

Proposition 3.10. (McDIARMID INEQUALITY) *Let $X = (X_1, \dots, X_n)$ where the X_i 's are independent r.v. with respected values in A_i . Let $f : \prod A_k \rightarrow \mathbb{R}$ verifying the following Lipschitz condition.*

$$\text{For any } x, x' \in \prod_{k=1}^n A_k, \quad |f(x) - f(x')| \leq c_k \quad \text{if } x_j = x'_j, \text{ for } j \neq k, \quad 1 \leq j \leq n. \quad (3.1)$$

Let us denote $\mu = \mathbb{E}[f(X)]$. Then, for any $t \geq 0$,

$$\mathbb{P}[f(X) - \mu \geq t] \leq e^{-2t^2 / \sum c_k^2}.$$

Moreover, the same inequality holds when replacing $f(X) - \mu$ by $\mu - f(X)$.

Proof. Lipschitz condition (3.1) implies that f is bounded, thus from Theorem 3.4 we have

$$\mathbb{P}[f(X) - \mu \geq t] \leq e^{-2t^2 / \hat{r}^2},$$

where \hat{r}^2 is defined by setting $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ and $X = f(X_1, \dots, X_n)$. Note that this inequality holds true only under the assumption that f is bounded, without independence assumption or Lipschitz condition. The latter two allows to derive an upper bound on \hat{r}^2 : $\text{ran}_k = \text{ran}(\mathbb{E}(f(X)|\mathcal{F}_k) - \mathbb{E}[f(X)|\mathcal{F}_{k-1}] | \mathcal{F}_{k-1}) \leq c_k$. \square

Proposition 3.11. (HOEFFDING INEQUALITY) *Let X_1, \dots, X_n be n independent random variables such that $a_i \leq X_i \leq b_i$, $1 \leq i \leq n$. Define $S_n = \sum X_k$ and $\mu = \mathbb{E}(S_n)$. Then,*

$$\mathbb{P}(S_n - \mu \geq t) \leq e^{-2t^2 / \sum (b_k - a_k)^2}.$$

Moreover, the same inequality holds when replacing $S_n - \mu$ by $\mu - S_n$.

Proof. This is a immediate consequence of previous McDiarmid inequality (Proposition 3.10) with $A_k = [a_k, b_k]$, $f(x) = \sum x_k$ and $c_k = b_k - a_k$. Within this setting, $\hat{r}^2 \leq b_k - a_k$. \square

Remark 3.12. This result can be directly proved with the classical lines as in Theorem 3.4: Exponential Markov inequality, sum of independent variable assumption (or of martingale differences), and use of Lemma 3.6 before optimization in h :

$$\begin{aligned} \mathbb{P}(S_n - \mu \geq t) &\leq \mathbb{E}(e^{h(S_n - \mu)})e^{-ht} \\ \mathbb{E}\left(\prod e^{h(X_k - \mathbb{E}X_k)}\right) &= \prod \mathbb{E}(e^{h(X_k - \mathbb{E}X_k)}) \quad (\text{from independence}) \\ &\leq e^{\frac{1}{8}h^2 \sum (b_k - a_k)^2} \quad (\text{from Lemma 3.6}), \end{aligned}$$

then setting $h = \frac{4t}{\sum (b_k - a_k)^2}$.

Remark 3.13. Comparing the two previous McDiarmid and Hoeffding inequalities with Theorem 3.4, we can appreciate that martingale differences decomposition allows to generalize the case of a sum of independent r.v. . Subject to introducing more precise control tools like \hat{r}^2 , independence or Lipschitz condition are not needed anymore. The two latter additional assumptions simply allows to bound \hat{r}^2 .

The three previous propositions ignore information about the variance of the underlying process. The following inequality deriving from Theorem 3.7 provides an improvement in this respect.

Proposition 3.14. (BERNSTEIN INEQUALITY) *Let X_1, \dots, X_n be n independent random variables with $X_k - \mathbb{E}(X_k) \leq b$. We consider their sum $S_n = \sum X_k$, the sum variance $V = \text{var}(S_n)$ as well as the sum expectation $\mathbb{E}(S_n) = \mu$. Then, for any $t \geq 0$,*

$$\mathbb{P}(S_n - \mu \geq t) \leq e^{-\frac{t^2}{2(V+bt/3)}},$$

and more generally,

$$\mathbb{P}((S_n - \mu \geq t) \cap (V \leq v)) \leq e^{-\frac{t^2}{2(v+bt/3)}}.$$

Remark 3.15. (GAIN WITH RESPECT TO INEQUALITIES WITHOUT VARIANCE TERM) Assume $0 \leq X_i \leq 1$ and consider renormalized quantities, namely $\tilde{S}_n := S_n/n$, $\tilde{\mu} := \mu/n$, $\tilde{V} = V/n^2$. Then,

$$\mathbb{P}(\tilde{S}_n - \tilde{\mu} \geq t) \leq e^{-2nt^2} \quad (\text{Hoeffding})$$

$$\text{and } \mathbb{P}(\tilde{S}_n - \tilde{\mu} \geq t) \leq e^{-\frac{nt^2}{2(\tilde{V}+t/3)}} \quad (\text{Bernstein}),$$

with t typically of order between $1/n$ and $1/\sqrt{n}$. Thus, if the variance \tilde{V} is small enough, Bernstein inequality allows to have rates in e^{-nt} instead of e^{-nt^2} . In other words, Bernstein-type inequality may give high probability bounds in $\frac{1}{n} \log \frac{1}{\delta}$ instead of $\sqrt{\frac{1}{n} \log \frac{1}{\delta}}$. This fact will be a keystone for deriving concentration bounds on low probability regions.

Proof. Let $F_k = \sigma(X_1, \dots, X_n)$, $X = \sum (X_k - \mathbb{E}X_k) = S_n - \mu$, $\tilde{X}_k = \mathbb{E}(X|\mathcal{F}_k) = \sum_{i=1}^k (X_i - \mathbb{E}X_i)$ and $Y_k = \tilde{X}_k - \tilde{X}_{k-1}$. Then $Y_k = X_k - \mathbb{E}X_k$, hence $\text{dev}_k^+ \leq b$, $\text{maxdev}^+ \leq b$ and $\text{var}_k = \text{var}(Y_k|\mathcal{F}_{k-1}) = \mathbb{E}((Y_k - \mathbb{E}(Y_k|\mathcal{F}_{k-1}))^2|\mathcal{F}_{k-1}) = \mathbb{E}((Y_k - \mathbb{E}Y_k)^2) = \text{var}(Y_k)$. Therefore $\hat{v} = \sup(\sum \text{var}_k) = \sup(V) = V$. Theorem 3.7 applies and yields,

$$\mathbb{P}(S_n - \mu \geq t) \leq e^{-\frac{t^2}{2(V+bt/3)}},$$

$$\mathbb{P}((S_n - \mu \geq t) \cap (V \leq v)) \leq e^{-\frac{t^2}{2(v+bt/3)}}.$$

□

3.3 Links with Statistical Learning and VC theory

In statistical learning theory, we are often interested in deriving concentration inequalities for the random variable

$$f(\mathbf{X}_1, \dots, \mathbf{X}_n) = \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right|,$$

where $\mathbf{X}_1, \dots, \mathbf{X}_n$ are *i.i.d.* realizations of a r.v. \mathbf{X} with values in \mathbb{R}^d and \mathcal{A} a class of subsets of \mathbb{R}^d . The class \mathcal{A} should be complex enough to provide small bias in the estimation process, while simple enough to provide small variance (avoiding over-fitting). Typically, \mathcal{A} will be a so-called *VC-class*, meaning that the following *VC-shatter coefficient*,

$$S_{\mathcal{A}}(n) = \max_{x_1, \dots, x_d \in \mathbb{R}^d} |\{\{x_1, \dots, x_n\} \cap A, A \in \mathcal{A}\}|, \quad (3.2)$$

can be bounded in that way,

$$S_{\mathcal{A}}(n) \leq (n+1)^{V_{\mathcal{A}}}, \quad (3.3)$$

where $V_{\mathcal{A}}$ is the VC-dimension of \mathcal{A} . $S_{\mathcal{A}}(n)$ is the maximal number of different subsets of a set of n points which can be obtained by intersecting it with elements of \mathcal{A} . Note that for any n , $S_{\mathcal{A}}(n) \leq 2^n$. For a very large class \mathcal{A} (those of infinite VC-dimension), we have $S_{\mathcal{A}}(n) = 2^n$ for all n . The VC-dimension of a class \mathcal{A} is precisely the larger number N such that $S_{\mathcal{A}}(N) = 2^N$. In that case, for $n \leq N$, $S_{\mathcal{A}}(n) = 2^n$.

As the variance of the r.v. $f(\mathbf{X}_1, \dots, \mathbf{X}_n)$ seems inaccessible, it is natural to apply a concentration inequality without variance term. It is easy to check that the function f verifies the Lipschitz condition 3.1 in McDiarmid inequality (Proposition 3.10), with $c_k = 1/n$. Thus, Proposition 3.10 yields

$$\mathbb{P}[f(\mathbf{X}_1, \dots, \mathbf{X}_n) - \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n) \geq t] \leq e^{-2nt^2},$$

or equivalently

$$f(\mathbf{X}_1, \dots, \mathbf{X}_n) \leq \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} \quad (3.4)$$

with probability at least $1 - \delta$. The complexity of class \mathcal{A} comes into play for bounding the expectation of $f(\mathbf{X}_1, \dots, \mathbf{X}_n)$. Consider the Rademacher average

$$\mathcal{R}_n = \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right|$$

where $(\sigma_i)_{i \geq 1}$ is a Rademacher chaos independent of the \mathbf{X}_i 's, namely the σ_i 's are *i.i.d.* with $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$. Then, the following result holds true.

Theorem 3.16. Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. random variables, and a VC-class \mathcal{A} with VC-dimension $V_{\mathcal{A}}$. The following inequalities hold true:

$$(i) \quad \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n) \leq 2\mathcal{R}_n$$

$$(ii) \quad \mathcal{R}_n \leq C\sqrt{\frac{V_{\mathcal{A}}}{n}}$$

Remark 3.17. Note that bound (ii) holds even for the conditional Rademacher average

$$\mathbb{E} \left[\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right| \mid \mathbf{X}_1, \dots, \mathbf{X}_n \right].$$

Proof. The second inequality is quite difficult to obtain and will not be detailed here. The proof of the second point is classical and relies on a *symetrization* step with a ghost sample \mathbf{X}'_i and a *randomization* step with a Rademacher chaos: Let $(\mathbf{X}'_i)_{1 \leq i \leq n}$ a ghost sample, namely i.i.d. independent copy of the \mathbf{X}_i 's, we may write:

$$\begin{aligned} \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n) &= \mathbb{E} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| \\ &= \mathbb{E} \sup_{A \in \mathcal{A}} \left| \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}'_i \in A} \right] - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| \\ &= \mathbb{E} \sup_{A \in \mathcal{A}} \left| \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}'_i \in A} - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \mid \mathbf{X}_1, \dots, \mathbf{X}_n \right] \right| \\ &\leq \mathbb{E} \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}'_i \in A} - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| \quad (\text{since } \mathbb{E} \sup(\cdot) \geq \sup \mathbb{E}(\cdot)) \\ &= \mathbb{E} \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i (\mathbb{1}_{\mathbf{X}'_i \in A} - \mathbb{1}_{\mathbf{X}_i \in A}) \right| \quad (\mathbb{1}_{\mathbf{X}'_i \in A} - \mathbb{1}_{\mathbf{X}_i \in A} \stackrel{\mathcal{L}}{=} \sigma_i (\mathbb{1}_{\mathbf{X}'_i \in A} - \mathbb{1}_{\mathbf{X}_i \in A})) \\ &\leq \mathbb{E} \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}'_i \in A} \right| + \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^n -\sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right| \\ &= 2\mathcal{R}_n \end{aligned}$$

□

Thus, combining Theorem 3.16 with (3.4) we obtain the following version of the famous Vapnik-Chervonenkis inequality.

Theorem 3.18. ((VAPNIK-CHERVONENKIS)) Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. random variables, and a VC-class \mathcal{A} with VC-dimension $V_{\mathcal{A}}$. For $\delta > 0$, with probability higher than $1 - \delta$:

$$f(\mathbf{X}_1, \dots, \mathbf{X}_n) \leq C\sqrt{\frac{V_{\mathcal{A}} + \log \frac{1}{\delta}}{n}}$$

In the literature, one often find a version of Vapnik-Chervonenkis inequality with an additional $\log n$ factor,

$$f(\mathbf{X}_1, \dots, \mathbf{X}_n) \leq C \sqrt{\frac{V_{\mathcal{A}} \log(n) + \log \frac{1}{\delta}}{n}}.$$

This can be explained by the fact that it is easy to prove the sub-optimal inequality $\mathcal{R}_n \leq C \sqrt{\frac{V_{\mathcal{A}} \log(n)}{n}}$.

When \mathcal{A} is finite with cardinal A , one can show that $\mathcal{R}_n \leq C \sqrt{\frac{A \log(n)}{n}}$.

3.4 Sharper VC-bounds through a Bernstein-type inequality

Contribution presented in Chapter 8 include some VC-type inequality obtained by using a Bernstein-type inequality instead of the McDiarmid one used in (3.4). As mentioned above, the variance of $f(\mathbf{X}_1, \dots, \mathbf{X}_n)$ seems inaccessible. For this reason, we have to consider more complex control tools like the maximum sum of conditional variances and apply the strong fundamental Theorem 3.7. The following lemma guarantees that the latter applies.

Lemma 3.19. *Consider the r.v. $f(\mathbf{X}_1, \dots, \mathbf{X}_n)$ defined above, and \maxdev^+ and \hat{v} respectively its associated maximum conditional deviation and associated maximum sum of conditional variances, both of which we assume to be finite. In this context,*

$$\maxdev^+ \leq \frac{1}{n} \text{ and } \hat{v} \leq \frac{q}{n}$$

where

$$q = \mathbb{E} \left(\sup_{A \in \mathcal{A}} \left| \mathbb{1}_{\mathbf{X}'_1 \in A} - \mathbb{1}_{\mathbf{X}_1 \in A} \right| \right) \leq 2 \mathbb{E} \left(\sup_{A \in \mathcal{A}} \left| \mathbb{1}_{\mathbf{X}'_1 \in A} \mathbb{1}_{\mathbf{X}_1 \notin A} \right| \right)$$

with \mathbf{X}'_1 an independent copy of \mathbf{X}_1 .

Proof. Introduce the functional

$$h(\mathbf{x}_1, \dots, \mathbf{x}_k) = \mathbb{E} [f(\mathbf{X}_1, \dots, \mathbf{X}_n) | \mathbf{X}_1 = \mathbf{x}_1, \dots, \mathbf{X}_k = \mathbf{x}_k] - \mathbb{E} [f(\mathbf{X}_1, \dots, \mathbf{X}_n) | \mathbf{X}_1 = \mathbf{x}_1, \dots, \mathbf{X}_{k-1} = \mathbf{x}_{k-1}]$$

The *positive deviation* of $h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{X}_k)$ is defined by

$$dev^+(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \{h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{x})\},$$

and \maxdev^+ , the maximum of all positive deviations, by

$$\maxdev^+ = \sup_{\mathbf{x}_1, \dots, \mathbf{x}_{k-1}} \max_k dev^+(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}).$$

Finally, \hat{v} , the *maximum sum of variances*, is defined by

$$\hat{v} = \sup_{\mathbf{x}_1, \dots, \mathbf{x}_n} \sum_{k=1}^n \mathbf{Var} \, h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{X}_k) .$$

Considering the definition of f , we have:

$$\begin{aligned} h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{x}_k) &= \mathbb{E} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^k \mathbb{1}_{\mathbf{x}_i \in A} - \frac{1}{n} \sum_{i=k+1}^n \mathbb{1}_{\mathbf{x}_i \in A} \right| \\ &\quad - \mathbb{E} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^{k-1} \mathbb{1}_{\mathbf{x}_i \in A} - \frac{1}{n} \sum_{i=k}^n \mathbb{1}_{\mathbf{x}_i \in A} \right| . \end{aligned}$$

Using the fact that $|\sup_{A \in \mathcal{A}} |F(A)| - \sup_{A \in \mathcal{A}} |G(A)|| \leq \sup_{A \in \mathcal{A}} |F(A) - G(A)|$ for every function F and G of A , we obtain:

$$|h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{x}_k)| \leq \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{n} |\mathbb{1}_{\mathbf{x}_k \in A} - \mathbb{1}_{\mathbf{x}_k \in A}| . \quad (3.5)$$

The term on the right hand side of (3.5) is less than $\frac{1}{n}$ so that $\max_{\text{dev}}^+ \leq \frac{1}{n}$. Moreover, if \mathbf{X}' is an independent copy of \mathbf{X} , (3.5) yields

$$|h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{X}')| \leq \mathbb{E} \left[\sup_{A \in \mathcal{A}} \frac{1}{n} |\mathbb{1}_{\mathbf{X}' \in A} - \mathbb{1}_{\mathbf{X} \in A}| \mid \mathbf{X}' \right] ,$$

so that

$$\begin{aligned} \mathbb{E} [h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{X}')^2] &\leq \mathbb{E} \mathbb{E} \left[\sup_{A \in \mathcal{A}} \frac{1}{n} |\mathbb{1}_{\mathbf{X}' \in A} - \mathbb{1}_{\mathbf{X} \in A}| \mid \mathbf{X}' \right]^2 \\ &\leq \mathbb{E} \left[\sup_{A \in \mathcal{A}} \frac{1}{n^2} |\mathbb{1}_{\mathbf{X}' \in A} - \mathbb{1}_{\mathbf{X} \in A}|^2 \right] \\ &\leq \frac{1}{n^2} \mathbb{E} \left[\sup_{A \in \mathcal{A}} |\mathbb{1}_{\mathbf{X}' \in A} - \mathbb{1}_{\mathbf{X} \in A}| \right] \end{aligned}$$

Thus $\mathbf{Var}(h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{X}_k)) \leq \mathbb{E}[h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{X}_k)^2] \leq \frac{q}{n^2}$. Finally $\hat{v} \leq \frac{q}{n}$ as required. \square

Remark 3.20. (ON PARAMETER q) The quantity $q = \mathbb{E}(\sup_{A \in \mathcal{A}} |\mathbb{1}_{\mathbf{X}' \in A} - \mathbb{1}_{\mathbf{X} \in A}|)$ is as a measure of the complexity of the class \mathcal{A} **with respect to the distribution of \mathbf{X}** (\mathbf{X}' being an independent copy of \mathbf{X}). It resembles to the Rademacher complexity \mathcal{R}_n . However, note that the latter is bounded *independently of the distribution of \mathbf{X}* , as bound (ii) in Theorem 3.16 holds for the conditional Rademacher average, namely for any distribution. Also note that $q \leq \sup_{A \in \mathcal{A}} \mathbb{P}(\mathbf{X} \in A) \leq \mathbb{P}(\mathbf{X} \in \cup_{A \in \mathcal{A}} A := p)$, the probability of hitting the class \mathcal{A} at all.

Applying Theorem 3.7 we get

$$\mathbb{P}[f(\mathbf{X}_1, \dots, \mathbf{X}_n) - \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n) \geq t] \leq e^{-\frac{nt^2}{2q + \frac{2t}{3}}},$$

or equivalently

$$\mathbb{P}\left[\frac{1}{q}(f(\mathbf{X}_1, \dots, \mathbf{X}_n) - \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n)) \geq t\right] \leq e^{-\frac{nqt^2}{2 + \frac{2t}{3}}}.$$

Solving $\exp\left[-\frac{nqt^2}{4 + \frac{2}{3}t}\right] = \delta$ with $t > 0$ leads to

$$t = \frac{1}{3nq} \log \frac{1}{\delta} + \sqrt{\left(\frac{1}{3nq} \log \frac{1}{\delta}\right)^2 + \frac{4}{nq} \log \frac{1}{\delta}} := h(\delta)$$

so that

$$\mathbb{P}\left[\frac{1}{q}(f(\mathbf{X}_1, \dots, \mathbf{X}_n) - \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n)) > h(\delta)\right] \leq \delta$$

Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ if $a, b \geq 0$, we have $h(\delta) < \frac{2}{3nq} \log \frac{1}{\delta} + 2\sqrt{\frac{1}{nq} \log \frac{1}{\delta}}$ in such a way that, with probability at least $1 - \delta$

$$f(\mathbf{X}_1, \dots, \mathbf{X}_n) \leq \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n) + \frac{2}{3n} \log \frac{1}{\delta} + 2\sqrt{\frac{q}{2n} \log \frac{1}{\delta}} \quad (3.6)$$

instead of (3.4).

Remark 3.21. (EXPECTATION BOUND) By classical arguments (see the proof of Theorem 3.16 above), $\mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n) \leq q_n := \mathbb{E} \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \left(\mathbb{1}_{\mathbf{X}'_i \in A} - \mathbb{1}_{\mathbf{X}_i \in A} \right) \right|$. Using Massart's finite class Lemma, see Massart (2000), to show that $q_n \leq \sqrt{q} \sqrt{\frac{2V_{\mathcal{A}} \log(en/V_{\mathcal{A}})}{n}}$ yields

$$f(\mathbf{X}_1, \dots, \mathbf{X}_n) \leq \sqrt{q} \sqrt{\frac{12 \log \frac{1}{\delta} + 4V_{\mathcal{A}} \log(\frac{en}{V_{\mathcal{A}}})}{n}} + \frac{2}{3n} \log \frac{1}{\delta}$$

with probability at least $1 - \delta$.

Contributions detailed in Chapter 8 use the improved VC-inequality (3.6) with $p := \mathbb{P}(\mathbf{X} \in \cup_{A \in \mathcal{A}} A)$ instead of q (see Lemma 3.22), combined with the following lemma which slightly improves Theorem 3.16:

Lemma 3.22. *Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. random variables with values in \mathbb{R}^d as above, and a VC-class \mathcal{A} with VC-dimension $V_{\mathcal{A}}$. Recall that p is the probability of hitting the class at all $p =$*

$\mathbb{P}(\mathbf{X} \in \cup_{A \in \mathcal{A}} A)$. The following inequality holds true:

$$(i) \quad \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n) \leq 2\mathcal{R}_n$$

$$(ii') \quad \mathcal{R}_n \leq C \sqrt{\frac{pV_{\mathcal{A}}}{n}}$$

Proof. Denote by $\mathcal{R}_{n,p}$ the associated relative Rademacher average defined by

$$\mathcal{R}_{n,p} = \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right|.$$

Let us defined *i.i.d.* r.v. \mathbf{Y}_i independent from \mathbf{X}_i whose law is the law of \mathbf{X} conditioned on the event $\mathbf{X} \in \mathbb{A}$. It is easy to show that $\sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \stackrel{d}{=} \sum_{i=1}^{\kappa} \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A}$, where $\kappa \sim \text{Bin}(n, p)$ independent of the \mathbf{Y}_i 's. Thus,

$$\begin{aligned} \mathcal{R}_{n,p} &= \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right| = \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^{\kappa} \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A} \right| \\ &= \mathbb{E} \left[\mathbb{E} \left[\sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^{\kappa} \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A} \right| \mid \kappa \right] \right] \\ &= \mathbb{E} [\Phi(\kappa)] \end{aligned}$$

where

$$\phi(K) = \mathbb{E} \left[\sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^K \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A} \right| \right] = \frac{K}{np} \mathcal{R}_K \leq \frac{K}{np} \frac{C\sqrt{V_{\mathcal{A}}}}{\sqrt{K}}.$$

Thus,

$$\mathcal{R}_{n,p} \leq \mathbb{E} \left[\frac{\sqrt{\kappa}}{np} C \sqrt{V_{\mathcal{A}}} \right] \leq \frac{\sqrt{\mathbb{E}[\kappa]}}{np} C \sqrt{V_{\mathcal{A}}} \leq \frac{C\sqrt{V_{\mathcal{A}}}}{\sqrt{np}}.$$

Finally, $\mathcal{R}_n = p\mathcal{R}_{n,p} \leq C\sqrt{\frac{pV_{\mathcal{A}}}{n}}$ as required. \square

CHAPTER 4

Extreme Value Theory

Chapter abstract In this chapter, we provide a concise background on Extreme Value Theory (EVT). The tools needed to approach chapters 8 and 9 are introduced.

There are many books introducing extreme value theory, like Leadbetter et al. (1983), Resnick (1987), Coles (2001), Beirlant et al. (2004), de Haan & Ferreira (2006), and Resnick (2007). Our favorites are Resnick (2007) for its comprehensiveness while remaining accessible, and Coles (2001) for the emphasis it puts on intuition. For a focus on Multivariate Extremes, we recommend Chap.6 of Resnick (2007) (and in particular, comprehensive Thm.6.1, 6.2 and 6.3) completed with Chap.8 of Coles (2001) for additional intuition. For the hurried reader, the combination of Segers (2012b), the two introductory parts of Einmahl et al. (2012) and the first four pages of Coles & Tawn (1991) provides a quick but in-depth introduction to multivariate extremes value theory, and have been of precious help to the author.

Extreme Value Theory (EVT) develops models for learning the unusual rather than the usual, in order to provide a reasonable assessment of the probability of occurrence of rare events. Such models are widely used in fields involving risk management such as Finance, Insurance, Operation Research, Telecommunication or Environmental Sciences for instance. For clarity, we start off with recalling some key notions pertaining to (multivariate) EVT, that shall be involved in the formulation of the problem next stated and in its subsequent analysis.

Notation reminder Throughout this chapter and all along this thesis, bold symbols refer to multivariate quantities, and for $m \in \mathbb{R} \cup \{\infty\}$, \mathbf{m} denotes the vector (m, \dots, m) . Also, comparison operators between two vectors (or between a vector and a real number) are understood component-wise, *i.e.* ' $\mathbf{x} \leq \mathbf{z}$ ' means ' $x_j \leq z_j$ for all $1 \leq j \leq d$ ' and for any real number T , ' $\mathbf{x} \leq T$ ' means ' $x_j \leq T$ for all $1 \leq j \leq d$ '. We denote by $\lfloor u \rfloor$ the integer part of any real number u , by $u_+ = \max(0, u)$ its positive part and by $\delta_{\mathbf{a}}$ the Dirac mass at any point $\mathbf{a} \in \mathbb{R}^d$. For uni-dimensional random variables Y_1, \dots, Y_n , $Y_{(1)} \leq \dots \leq Y_{(n)}$ denote their order statistics.

4.1 Univariate Extreme Value Theory

In the univariate case, EVT essentially consists in modeling the distribution of the maxima (*resp.* the upper tail of the r.v. under study) as a *generalized extreme value distribution*, namely an element of the Gumbel, Fréchet or Weibull parametric families (*resp.* by a generalized Pareto distribution).

A useful setting to understand the use of EVT is that of risk monitoring. A typical quantity of interest in the univariate case is the $(1 - p)^{th}$ quantile of the distribution F of a r.v. X , for a given exceedance probability p , that is $x_p = \inf\{x \in \mathbb{R}, \mathbb{P}(X > x) \leq p\}$. For moderate values of p , a natural empirical estimate is

$$x_{p,n} = \inf\{x \in \mathbb{R}, 1/n \sum_{i=1}^n \mathbb{1}_{\{X_i > x\}} \leq p\}.$$

However, if p is very small, the finite sample X_1, \dots, X_n carries insufficient information and the empirical quantile $x_{p,n}$ becomes unreliable. That is where EVT comes into play by providing parametric estimates of large quantiles: whereas statistical inference often involves sample means and the Central Limit Theorem, EVT handles phenomena whose behavior is not ruled by an ‘averaging effect’. The focus is on the sample maximum

$$M_n = \max\{X_1, \dots, X_n\}$$

rather than the mean. A first natural approach is to estimate F from observed data to deduce an estimate of F^n , given that the distribution of M_n is

$$\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x) \dots \mathbb{P}(X_n \leq x) = F(x)^n.$$

Unfortunately, the exponent in the number of data induces huge discrepancies for such plug-in techniques. The next natural approach is to look directly for appropriate families of models for F^n . A first difficulty is that any point less than the upper point of F is finally exceeded by the maximum of a sufficiently large number of data: $F^n(x) \rightarrow 0$ for any x such that $F(x) < 1$. In other words, the distribution of M_n converge to a dirac mass on $\inf\{x, F(x) = 1\}$. Therefore, we have to consider a renormalized version of M_n ,

$$\frac{M_n - b_n}{a_n}$$

with $a_n > 0$. Then, the cornerstone result of univariate EVT is the following.

Theorem 4.1. *Assume there exists such sequences $\{a_n, n \geq 1\}$ and $\{b_n, n \geq 1\}$, the a_n ’s being positive, such that $\frac{M_n - b_n}{a_n}$ converges in distribution to a non-degenerate distribution, namely*

$$\mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = F^n(a_n x + b_n) \rightarrow G(x) \quad (4.1)$$

for all continuity point of G , where G is a non-degenerate distribution function (i.e. without dirac mass). Then G belongs to one of the three following extreme value distribution (up to a rescaling $x' = \frac{x-b}{a}$ which can be removed by changing a_n and b_n):

$$\text{Gumbel: } G(x) = \exp(-e^{-x}) \quad \text{for } x \in (-\infty, +\infty),$$

$$\text{Fréchet: } G(x) = \exp(-x^{-\alpha}) \quad \text{if } x > 0 \text{ and } G(x) = 0 \text{ otherwise,}$$

$$\text{Weibull: } G(x) = \exp(-(-x)^\alpha) \quad \text{if } x < 0 \text{ and } G(x) = 1 \text{ otherwise,}$$

with $\alpha > 0$.

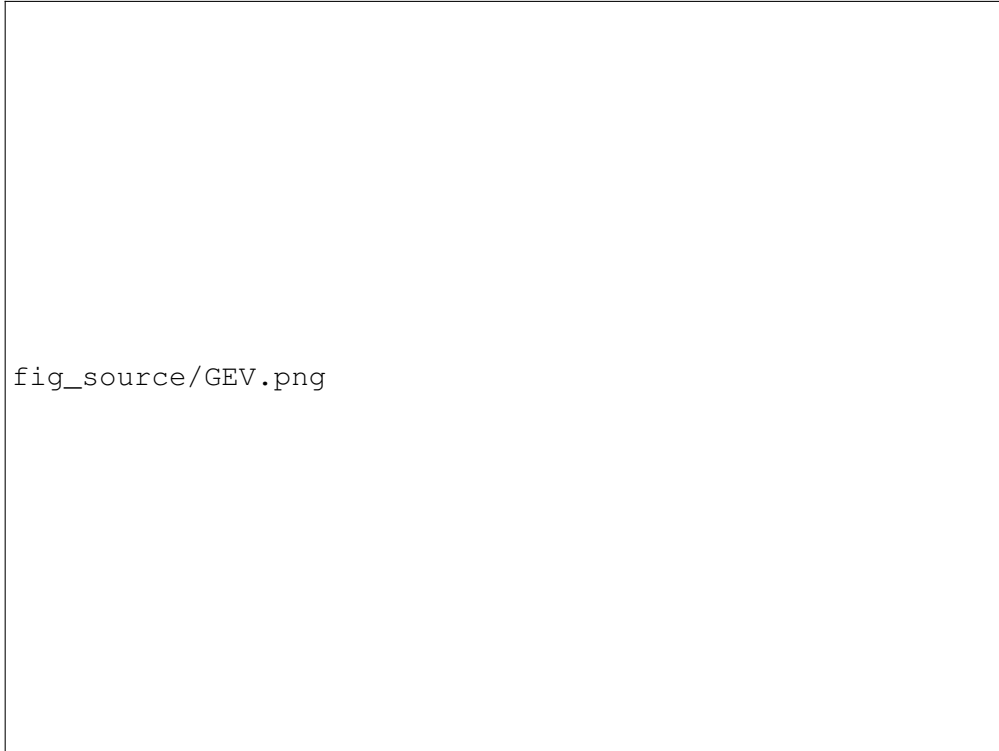


FIGURE 4.1: Extreme Value Distribution with $\alpha = 2$

These extreme value distributions plotted in Figure 4.1 can be summarized into the so-called Generalized Extreme Value (GEV) Distribution,

$$G(x) = \exp\left(-[1 + \gamma x]^{-1/\gamma}\right) \quad (4.2)$$

for $1 + \gamma x > 0$, $\gamma \in \mathbb{R}$, setting by convention $(1 + \gamma x)^{-1/\gamma} = e^{-x}$ for $\gamma = 0$ (continuous extension). The sign of γ controls the shape of the tail of F . In the case $\gamma > 0$ (as for the Cauchy distribution), G is referred to as a Fréchet distribution and F has a heavy tail. If $\gamma = 0$ (as for normal distributions), G is a Gumbel distribution and F has a light tail. If $\gamma < 0$ (as for uniform distributions), G is called a Weibull distribution and F has a finite endpoint. Estimates of univariate extreme quantiles then rely on estimates of the parameters a_n , b_n , and γ , see Dekkers et al. (1989), Einmahl et al. (2009). The Hill estimator or one of its generalizations, see Hill (1975), Smith (1987), Beirlant et al. (1996), provides an estimate of the tail parameter γ .

Example 4.1. • Assume the X_i 's to be standard exponential variables (their cdf is $F(x) = 1 - e^{-x}$). In that case, letting $a_n = 1$ and $b_n = \log(n)$, we have $\mathbb{P}[(M_n - b_n)/a_n \leq z] = \mathbb{P}[X_1 \leq z + \log n]^n = [1 - e^{-z}/n]^n \rightarrow \exp(-e^{-z})$, for $z \in \mathbb{R}$. The limit distribution is of Gumbel type ($\xi = 0$).

- If the X_i 's are standard Fréchet ($F(x) = \exp(-1/x)$), letting $a_n = n$ and $b_n = 0$, one has immediately $\mathbb{P}[(M_n - b_n)/a_n \leq z] = F^n(nz) = \exp(-1/z)$, for $z > 0$. The limit distribution remains the Fréchet one ($\xi = 1$).
- If the X_i 's are uniform on $[0, 1]$, letting $a_n = 1/n$ and $b_n = 1$, one has $\mathbb{P}[(M_n - b_n)/a_n \leq z] = F^n(n^{-1}z + 1) \rightarrow \exp(z)$, for $z < 0$. The limit distribution is the Weibull one ($\xi = -1$).

One can establish an equivalent formulation of Assumption (4.1) which does not rely anymore on the maximum M_n :

$$\lim_{n \rightarrow \infty} n \mathbb{P} \left(\frac{X - b_n}{a_n} \geq x \right) = -\log G(x) \quad (4.3)$$

for all continuity points $x \in \mathbb{R}$ of G . The intuition behind this equivalence is that

$$-\log(F^n(a_n x + b_n)) \sim n(1 - F(a_n x + b_n)) = n \mathbb{P} \left(\frac{X - b_n}{a_n} \geq x \right)$$

when $n \rightarrow \infty$ as $F(a_n x + b_n) \sim 1$. The tail behavior of F is then essentially characterized by G , which is proved to be – up to re-scaling – of the type (4.2). Note that Assumption (4.1) (or (4.3)) is fulfilled for most textbook distributions. In that case F is said to lie in the *domain of attraction* of G , written $F \in DA(G)$.

4.2 Extension to the Multivariate framework

Extensions to the multivariate setting are well understood from a probabilistic point of view, but far from obvious from a statistical perspective. Indeed, the tail dependence structure, ruling the possible simultaneous occurrence of large observations in several directions, has no finite-dimensional parametrization.

The analogue of Assumption (4.3) for a d -dimensional r.v. $\mathbf{X} = (X^1, \dots, X^d)$ with distribution $\mathbf{F}(\mathbf{x}) := \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d)$, written $\mathbf{F} \in \mathbf{DA}(\mathbf{G})$ stipulates the existence of two sequences $\{\mathbf{a}_n, n \geq 1\}$ and $\{\mathbf{b}_n, n \geq 1\}$ in \mathbb{R}^d , the \mathbf{a}_n 's being positive, and a non-degenerate distribution function \mathbf{G} such that

$$\lim_{n \rightarrow \infty} n \mathbb{P} \left(\frac{X^1 - b_n^1}{a_n^1} \geq x_1 \text{ or } \dots \text{ or } \frac{X^d - b_n^d}{a_n^d} \geq x_d \right) = -\log \mathbf{G}(\mathbf{x}) \quad (4.4)$$

for all continuity points $\mathbf{x} \in \mathbb{R}^d$ of \mathbf{G} . This clearly implies that the margins $G_1(x_1), \dots, G_d(x_d)$ are univariate extreme value distributions, namely of the type $G_j(x) = \exp(-(1 + \gamma_j x)^{-1/\gamma_j})$. Also, denoting by F_1, \dots, F_d the marginal distributions of \mathbf{F} , Assumption (4.4) implies marginal convergence: $F_i \in DA(G_i)$ for $i = 1, \dots, d$. To understand the structure of the limit \mathbf{G} and dispose of the unknown sequences $(\mathbf{a}_n, \mathbf{b}_n)$ (which are entirely determined by the marginal distributions F_j 's), it is convenient to work with marginally standardized variables, that is, to separate the margins from the dependence structure in the description of the joint distribution of \mathbf{X} . Consider the standardized variables $V^j = 1/(1 - F_j(X^j))$ and $\mathbf{V} = (V^1, \dots, V^d)$. In fact (see Proposition 5.10 in Resnick (1987)), Assumption (4.4) is equivalent to:

- marginal convergences $F_j \in DA(G_j)$ as in (4.3), together with
- standard multivariate regular variation of \mathbf{V} 's distribution, which means existence of a limit measure μ on $[0, \infty]^d \setminus \{\mathbf{0}\}$ such that

$$n \mathbb{P} \left(\frac{V^1}{n} \geq v_1 \text{ or } \dots \text{ or } \frac{V^d}{n} \geq v_d \right) \xrightarrow[n \rightarrow \infty]{} \mu([\mathbf{0}, \mathbf{v}]^c), \quad (4.5)$$

where $[\mathbf{0}, \mathbf{v}] := [0, v_1] \times \dots \times [0, v_d]$.

Thus the variable \mathbf{V} satisfies (4.4) with $\mathbf{a}_n = \mathbf{n} = (n, \dots, n)$, $\mathbf{b}_n = \mathbf{0} = (0, \dots, 0)$.

Remark 4.2. The standardization in V allows to study the same extreme value distribution for each marginal, and with the same re-scaling sequences a_n and b_n for each marginal. In the case of Pareto standardization like here, the underlying extreme value distribution is the Fréchet one.

The dependence structure of the limit \mathbf{G} in (4.4) can be expressed by means of the so-termed *exponent measure* μ :

$$-\log \mathbf{G}(\mathbf{x}) = \mu \left(\left[\mathbf{0}, \left(\frac{-1}{\log G_1(x_1)}, \dots, \frac{-1}{\log G_d(x_d)} \right) \right]^c \right).$$

The latter is finite on sets bounded away from $\mathbf{0}$ and has the homogeneity property : $\mu(t \cdot) = t^{-1} \mu(\cdot)$. Observe in addition that, due to the standardization chosen (with ‘nearly’ Pareto margins), the support of μ is included in $[\mathbf{0}, \mathbf{1}]^c$. To wit, the measure μ should be viewed, up to a normalizing factor, as the asymptotic distribution of \mathbf{V} in extreme regions. For any borelian subset A bounded away from $\mathbf{0}$ on which μ is continuous, we have

$$t \mathbb{P}(\mathbf{V} \in tA) \xrightarrow[t \rightarrow \infty]{} \mu(A). \quad (4.6)$$

Using the homogeneity property $\mu(t \cdot) = t^{-1} \mu(\cdot)$, one may show that μ can be decomposed into a radial component and an angular component Φ , which are independent from each other

(see *e.g.* de Haan & Resnick (1977)). Indeed, for all $\mathbf{v} = (v_1, \dots, v_d) \in \mathbb{R}^d$, set

$$\begin{cases} R(\mathbf{v}) := \|\mathbf{v}\|_\infty = \max_{i=1}^d v_i, \\ \Theta(\mathbf{v}) := \left(\frac{v_1}{R(\mathbf{v})}, \dots, \frac{v_d}{R(\mathbf{v})} \right) \in S_\infty^{d-1}, \end{cases} \quad (4.7)$$

where S_∞^{d-1} is the positive orthant of the unit sphere in \mathbb{R}^d for the infinity norm. Define the *spectral measure* (also called *angular measure*) by $\Phi(B) = \mu(\{\mathbf{v} : R(\mathbf{v}) > 1, \Theta(\mathbf{v}) \in B\})$. Then, for every $B \subset S_\infty^{d-1}$,

$$\mu\{\mathbf{v} : R(\mathbf{v}) > z, \Theta(\mathbf{v}) \in B\} = z^{-1} \Phi(B). \quad (4.8)$$

In a nutshell, there is a one-to-one correspondence between the exponent measure μ and the angular measure Φ , both of them can be used to characterize the asymptotic tail dependence of the distribution \mathbf{F} (as soon as the margins F_j are known), since

$$\mu([\mathbf{0}, \mathbf{x}^{-1}]^c) = \int_{\boldsymbol{\theta} \in S_\infty^{d-1}} \max_j \theta_j x_j \, d\Phi(\boldsymbol{\theta}), \quad (4.9)$$

this equality being obtained from the change of variable (4.7), see *e.g.* Proposition 5.11 in Resnick (1987). Recall that here and beyond, operators on vectors are understood component-wise, so that $\mathbf{x}^{-1} = (x_1^{-1}, \dots, x_d^{-1})$. The angular measure can be seen as the asymptotic conditional distribution of the ‘angle’ Θ given that the radius R is large, up to the normalizing constant $\Phi(S_\infty^{d-1})$. Indeed, dropping the dependence on \mathbf{V} for convenience, we have for any *continuity set* A of Φ ,

$$\mathbb{P}(\Theta \in A \mid R > r) = \frac{r\mathbb{P}(\Theta \in A, R > r)}{r\mathbb{P}(R > r)} \xrightarrow{r \rightarrow \infty} \frac{\Phi(A)}{\Phi(S_\infty^{d-1})}. \quad (4.10)$$

The choice of the marginal standardization is somewhat arbitrary and alternative standardizations lead to different limits. Another common choice consists in considering ‘nearly uniform’ variables (namely, uniform variables when the margins are continuous): defining \mathbf{U} by $U^j = 1 - F_j(X^j)$ for $j \in \{1, \dots, d\}$, condition (4.5) is equivalent to each of the following conditions:

- \mathbf{U} has ‘inverse multivariate regular variation’ with limit measure $\Lambda(\cdot) := \mu((\cdot)^{-1})$, namely, for every measurable set A bounded away from $+\infty$ which is a continuity set of Λ ,

$$t \mathbb{P}(\mathbf{U} \in t^{-1}A) \xrightarrow{t \rightarrow \infty} \Lambda(A) = \mu(A^{-1}), \quad (4.11)$$

where $A^{-1} = \{\mathbf{u} \in \mathbb{R}_+^d : (u_1^{-1}, \dots, u_d^{-1}) \in A\}$. The limit measure Λ is finite on sets bounded away from $\{+\infty\}$.

- The *stable tail dependence function* (STDF) defined for $\mathbf{x} \in [0, \infty]$, $\mathbf{x} \neq \infty$ by

$$l(\mathbf{x}) = \lim_{t \rightarrow 0} t^{-1} \mathbb{P} \left(U^1 \leq t x_1 \text{ or } \dots \text{ or } U^d \leq t x_d \right) = \mu \left([0, \mathbf{x}^{-1}]^c \right) \quad (4.12)$$

exists.

As a conclusion, in multivariate extremes, the focus is on the dependence structure which is characterized by different quantities, such as the exponent measure μ (itself characterized by its angular part Φ) or the STDF, which is closely linked to other integrated version of μ such as extreme-value copula or tail copula. For details on such functionals, see Segers (2012b). The fact that these quantities characterize the dependence structure can be illustrated by the link they exhibit between the multivariate GEV $G(x)$ and the marginal ones $G_j(x_j)$, $1 \leq j \leq d$,

$$\begin{aligned} -\log \mathbf{G}(\mathbf{x}) &= \mu \left(\left[0, \left(\frac{-1}{\log G_1(x_1)}, \dots, \frac{-1}{\log G_d(x_d)} \right) \right]^c \right) && \text{for the exponent measure,} \\ -\log \mathbf{G}(\mathbf{x}) &= l(-\log G_1(x_1), \dots, -\log G_d(x_d)) && \text{for the STDF,} \\ G(\mathbf{x}) &= C(G_1(x_1), \dots, G_d(x_d)) && \text{for the extreme value copula } C. \end{aligned}$$

In Chapter 8, we develop non-asymptotic bounds for non-parametric estimation of the STDF. As in many applications, it can be more convenient to work with the angular measure itself – the latter gives more direct information on the dependence structure –, Chapter 9 generalizes the study in Chapter 8 to the angular measure.

PART II

**An Excess-Mass based
Performance Criterion**

Chapter abstract This chapter presents the details relative to the introducing section 1.3. Learning how to rank multivariate unlabeled observations depending on their degree of abnormality/novelty is a crucial problem in a wide range of applications. In practice, it generally consists in building a real valued ‘scoring’ function on the feature space so as to quantify to which extent observations should be considered as abnormal. In the 1-d situation, measurements are generally considered as ‘abnormal’ when they are remote from central measures such as the mean or the median. Anomaly detection then relies on tail analysis of the variable of interest. Extensions to the multivariate setting are far from straightforward and it is precisely the main purpose of this chapter to introduce a novel and convenient (functional) criterion for measuring the performance of a scoring function regarding the anomaly ranking task, referred to as the *Excess-Mass* curve (*EM* curve). In addition, an adaptive algorithm for building a scoring function based on unlabeled data X_1, \dots, X_n with a nearly optimal *EM*-curve is proposed and is analyzed from a statistical perspective. The material of this chapter is based on previous work published in Goix et al. (2015c).

5.1 Introduction

In a great variety of applications (*e.g.* fraud detection, distributed fleet monitoring, system management in data centers), it is of crucial importance to address anomaly/novelty issues from a ranking point of view. In contrast to novelty/anomaly detection (*e.g.* Koltchinskii (1997); Vert & Vert (2006); Schölkopf et al. (2001); Steinwart et al. (2005)), novelty/anomaly ranking is very poorly documented in the statistical learning literature (see Viswanathan et al. (2012) for instance). However, when confronted with massive data, being able to rank observations according to their supposed degree of abnormality may significantly improve operational processes and allow for a prioritization of actions to be taken, especially in situations where human expertise required to check each observation is time-consuming. When univariate, observations are usually considered as ‘abnormal’ when they are either too high or else too small compared to central measures such as the mean or the median. In this context, anomaly/novelty analysis generally relies on the analysis of the tail distribution of the variable of interest. No natural (pre) order exists on a d -dimensional feature space, $\mathcal{X} \subset \mathbb{R}^d$ say, as soon as $d > 1$. Extension to the multivariate setup is thus far from obvious and, in practice, the optimal ordering/ranking

must be *learned* from training data X_1, \dots, X_n , in absence of any parametric assumptions on the underlying probability distribution describing the ‘normal’ regime. The most straightforward manner to define a preorder on the feature space \mathcal{X} is to transport the natural order on the real half-line through a measurable *scoring function* $s : \mathcal{X} \rightarrow \mathbb{R}_+$: the ‘smaller’ the score $s(X)$, the more ‘abnormal’ the observation X is viewed. In the following, to simplify notation we assume that $\mathcal{X} = \mathbb{R}^d$. The whys and wherefores of scoring functions have been explained in the introduction chapter, Section 1.2. Estimating good scoring functions is a way to estimate level sets of the underlying density, as optimal scoring function are those whose induced level sets are exactly the ones of the density. The basic idea is that we don’t need to estimate the density to obtain such level sets, but only any increasing transform of the density. Any scoring function defines a preorder on \mathbb{R}^d and thus a ranking on a set of new observations. An important issue stated in Section 1.2 concerns the definition of an adequate performance criterion, $\mathcal{C}(s)$ say, in order to compare possible candidate scoring function and to pick one eventually: optimal scoring functions s^* being then defined as those optimizing \mathcal{C} . Estimating scoring function instead of the density itself precisely allows to use an other criterion than the distance to the density, which is too stringent for a level sets estimation purpose as function having exactly the same level sets as the density can be very far from the latter which such distance.

Throughout the present article, it is assumed that the distribution F of the observable r.v. X is absolutely continuous w.r.t. Lebesgue measure Leb on \mathbb{R}^d , with density $f(x)$. The criterion should be thus defined in a way that the collection of level sets of an optimal scoring function $s^*(x)$ coincides with that related to f . In other words, any nondecreasing transform of the density should be optimal regarding the ranking performance criterion \mathcal{C} . According to the Empirical Risk Minimization (ERM) paradigm, a scoring function will be built in practice by optimizing an empirical version $\mathcal{C}_n(s)$ of the criterion over an adequate set of scoring functions \mathcal{S}_0 of controlled complexity (e.g. a major class of finite VC dimension). Hence, another desirable property to guarantee the universal consistency of ERM learning strategies is the uniform convergence of $\mathcal{C}_n(s)$ to $\mathcal{C}(s)$ over such collections \mathcal{S}_0 under minimal assumptions on the distribution $F(dx)$.

As described in Section 1.3.2, a functional criterion referred to as the mass-volume curve (*MV-curve*), admissible with respect to the requirements listed above has been introduced in Cl  men  on & Jakubowicz (2013), extending somehow the concept of ROC curve in the unsupervised setup. Relying on the theory of *minimum volume* sets (see Section 1.3.1), it has been proved that the scoring functions minimizing empirical and discretized versions of the *MV-curve* criterion are accurate when the underlying distribution has compact support and a first algorithm for building nearly optimal scoring functions, based on the estimate of a finite collection of properly chosen minimum volume sets, has been introduced and analyzed. However, as explained in Section 1.3.2, some important drawbacks are inherent to this mass-volume curve criterion:

- 1) When used as an performance criterion, the lebesgue measure of possibly very complex sets has to be compute.

- 2) When used as an performance criterion, the pseudo-inverse $\alpha_s^{-1}(\alpha)$ may be hard to compute.
- 3) When used as a learning criterion (in the ERM paradigm), it produces level sets which are not necessarily nested, on which may be built inaccurate scoring function.
- 4) When used as a learning criterion, the learning rates are rather slow (of the order $n^{-1/4}$ namely), and cannot be established in the unbounded support situation.

Given these limitations, it is the major goal of this chapter to propose an alternative criterion for anomaly ranking/scoring, called the *Excess-Mass* curve (*EM* curve in short) here, based on the notion of *density contour clusters* Polonik (1995); Hartigan (1987); Müller & Sawitzki (1991). Whereas minimum volume sets are solutions of volume minimization problems under mass constraints, the latter are solutions of mass maximization under volume constraints. Exchanging this way objective and constraint, the relevance of this performance measure is thoroughly discussed and accuracy of solutions which optimize statistical counterparts of this criterion is investigated. More specifically, rate bounds of the order $n^{-1/2}$ are proved, even in the case of unbounded support. Additionally, in contrast to the analysis carried out in Cléménçon & Jakubowicz (2013), the model bias issue is tackled, insofar as the assumption that the level sets of the underlying density $f(x)$ belongs to the class of sets used to build the scoring function is relaxed here.

The rest of this chapter is organized as follows. Section 5.3 introduces the notion of *EM* curve and that of optimal *EM* curve. Estimation in the compact support case is covered by Section 5.4, extension to distributions with non compact support and control of the model bias are tackled in Section 5.5. A simulation study is performed in Section 5.6. All proofs are deferred to the last section 5.9.

5.2 Background and related work

As a first go, we first recall the *MV* curve criterion approach as introduced in Section 1.3.2, as a basis for comparison with that promoted in the present contribution.

Recall that \mathcal{S} is the set of all scoring functions $s : \mathbb{R}^d \rightarrow \mathbb{R}_+$ integrable w.r.t. Lebesgue measure. Let $s \in \mathcal{S}$. As defined in Cléménçon & Jakubowicz (2013); Cléménçon & Robbiano (2014), the *MV*-curve of s is the plot of the mapping

$$\alpha \in (0, 1) \mapsto MV_s(\alpha) = \lambda_s \circ \alpha_s^{-1}(\alpha),$$

where

$$\begin{aligned} \alpha_s(t) &= \mathbb{P}(s(X) \geq t), \\ \lambda_s(t) &= \text{Leb}(\{x \in \mathbb{R}^d, s(x) \geq t\}) \end{aligned} \tag{5.1}$$

and H^{-1} denotes the pseudo-inverse of any cdf $H : \mathbb{R} \rightarrow (0, 1)$. This induces a partial ordering on the set of all scoring functions: s is preferred to s' if $MV_s(\alpha) \leq MV_{s'}(\alpha)$ for all $\alpha \in (0, 1)$. One may show that $MV^*(\alpha) \leq MV_s(\alpha)$ for all $\alpha \in (0, 1)$ and any scoring function s , where $MV^*(\alpha)$ is the optimal value of the constrained minimization problem

$$\min_{\Gamma \text{ borelian}} \text{Leb}(\Gamma) \text{ subject to } \mathbb{P}(X \in \Gamma) \geq \alpha. \quad (5.2)$$

Suppose now that $F(dx)$ has a density $f(x)$ satisfying the following assumptions:

A₁ The density f is bounded, i.e. $\|f(X)\|_\infty < +\infty$.

A₂ The density f has no flat parts: $\forall c \geq 0, \mathbb{P}\{f(X) = c\} = 0$.

One may then show that the curve MV^* is actually a MV curve, that is related to (any increasing transform of) the density f namely: $MV^* = MV_f$. In addition, the minimization problem (5.2) has a unique solution Γ_α^* of mass α exactly, referred to as *minimum volume set* (see Section 1.3.1):

$$MV^*(\alpha) = \text{Leb}(\Gamma_\alpha^*) \text{ and } F(\Gamma_\alpha^*) = \alpha.$$

Anomaly scoring can be then viewed as the problem of building a scoring function $s(x)$ based on training data such that MV_s is (nearly) minimum everywhere, i.e. minimizing

$$\|MV_s - MV^*\|_\infty := \sup_{\alpha \in [0,1]} |MV_s(\alpha) - MV^*(\alpha)|.$$

Since F is unknown, a minimum volume set estimate $\hat{\Gamma}_\alpha^*$ can be defined as the solution of (5.2) when F is replaced by its empirical version $F_n = (1/n) \sum_{i=1}^n \delta_{X_i}$, minimization is restricted to a collection \mathcal{G} of borelian subsets of \mathbb{R}^d supposed not too complex but rich enough to include all density level sets (or reasonable approximants of the latter) and α is replaced by $\alpha - \phi_n$, where the *tolerance parameter* ϕ_n is a probabilistic upper bound for the supremum $\sup_{\Gamma \in \mathcal{G}} |F_n(\Gamma) - F(\Gamma)|$. Refer to Scott & Nowak (2006) for further details. The set \mathcal{G} should ideally offer statistical and computational advantages both at the same time. Allowing for fast search on the one hand and being sufficiently complex to capture the geometry of target density level sets on the other. In Cl  men  on & Jakubowicz (2013), a method consisting in preliminarily estimating a collection of minimum volume sets related to target masses $0 < \alpha_1 < \dots < \alpha_K < 1$ forming a subdivision of $(0, 1)$ based on training data so as to build a scoring function

$$s = \sum_k \mathbf{1}_{x \in \hat{\Gamma}_{\alpha_k}^*}$$

has been proposed and analyzed. Under adequate assumptions (related to \mathcal{G} , the perimeter of the $\Gamma_{\alpha_k}^*$'s and the subdivision step in particular) and for an appropriate choice of $K = K_n$ either under the very restrictive assumption that $F(dx)$ is compactly supported or else by restricting the convergence analysis to $[0, 1 - \epsilon]$ for $\epsilon > 0$, excluding thus the tail behavior of the distribution F from the scope of the analysis, rate bounds of the order $\mathcal{O}_{\mathbb{P}}(n^{-1/4})$ have been established to guarantee the generalization ability of the method.

Figure 5.3 illustrates one problem inherent to the use of the MV curve as a performance criterion for anomaly scoring in a ‘non asymptotic’ context, due to the prior discretization along the mass-axis. In the 2-d situation described by Figure 5.3 for instance, given the training sample and the partition of the feature space depicted, the MV criterion leads to consider the sequence of empirical minimum volume sets $A_1, A_1 \cup A_2, A_1 \cup A_3, A_1 \cup A_2 \cup A_3$ and thus the scoring function $s_1(x) = \mathbb{I}\{x \in A_1\} + \mathbb{I}\{x \in A_1 \cup A_2\} + \mathbb{I}\{x \in A_1 \cup A_3\}$, whereas the scoring function $s_2(x) = \mathbb{I}\{x \in A_1\} + \mathbb{I}\{x \in A_1 \cup A_3\}$ is clearly more accurate.

In this work, a different functional criterion is proposed, obtained by exchanging objective and constraint functions in (5.2), and it is shown that optimization of an empirical discretized version of this performance measure yields scoring rules with convergence rates of the order $\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})$. In addition, the results can be extended to the situation where the support of the distribution F is not compact.

5.3 The Excess-Mass curve

As introduced in Section 1.3.3, the performance criterion we propose in order to evaluate anomaly scoring accuracy relies on the notion of *excess mass* and *density contour clusters*, as introduced in the seminal contribution Polonik (1995). The main idea is to consider a Lagrangian formulation of a constrained minimization problem, obtained by exchanging constraint and objective in (5.2): for $t > 0$,

$$\max_{\Omega \text{ borelian}} \{ \mathbb{P}(X \in \Omega) - t \text{Leb}(\Omega) \}. \quad (5.3)$$

We denote by Ω_t^* any solution of this problem. As shall be seen in the subsequent analysis (see Proposition 5.6 below), compared to the MV curve approach, this formulation offers certain computational and theoretical advantages both at the same time: when letting (a discretized version of) the Lagrangian multiplier t increase from 0 to infinity, one may easily obtain solutions of empirical counterparts of (5.3) forming a *nested* sequence of subsets of the feature space, avoiding thus deteriorating rate bounds by transforming the empirical solutions so as to force monotonicity.

Definition 5.1. (OPTIMAL EM CURVE) The optimal Excess-Mass curve related to a given probability distribution $F(dx)$ is defined as the plot of the mapping

$$t > 0 \mapsto EM^*(t) := \max_{\Omega \text{ borelian}} \{ \mathbb{P}(X \in \Omega) - t \text{Leb}(\Omega) \}.$$

Equipped with the notation above, we have: $EM^*(t) = \mathbb{P}(X \in \Omega_t^*) - t \text{Leb}(\Omega_t^*)$ for all $t > 0$. Notice also that $EM^*(t) = 0$ for any $t > \|f\|_{\infty} := \sup_{x \in \mathbb{R}^d} |f(x)|$.

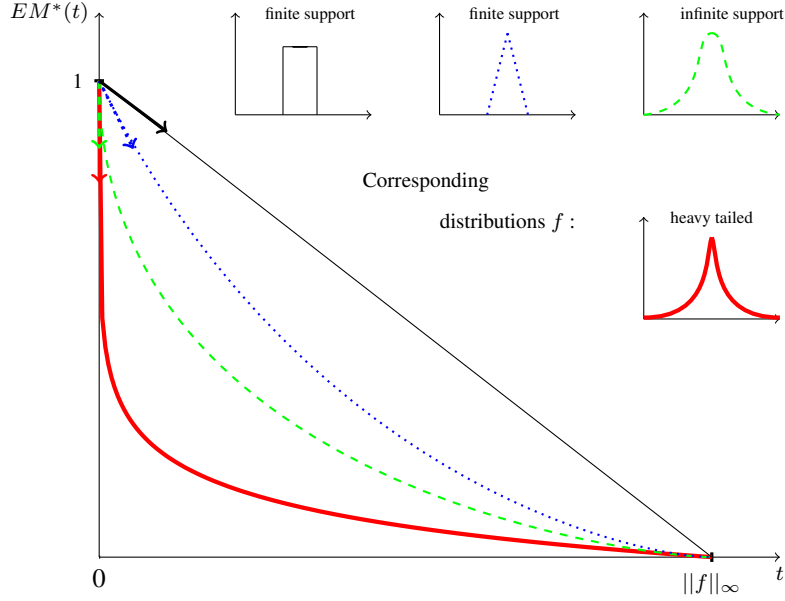
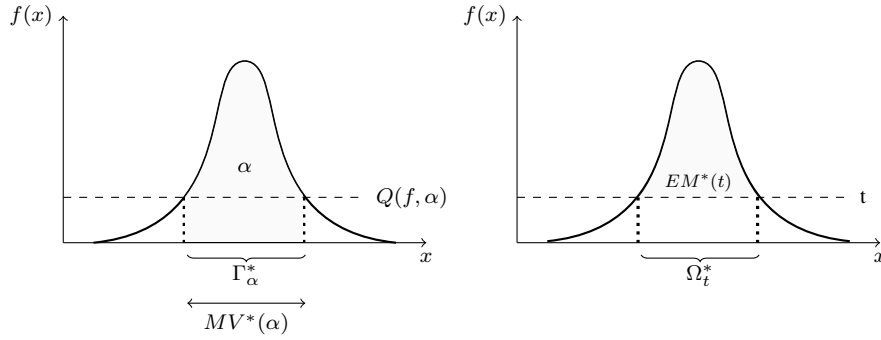


FIGURE 5.1: EM curves depending on densities

FIGURE 5.2: Comparison between $MV^*(\alpha)$ and $EM^*(t)$

Lemma 5.2. (ON EXISTENCE AND UNIQUENESS) *For any subset Ω_t^* solution of (5.3), we have*

$$\{x, f(x) > t\} \subset \Omega_t^* \subset \{x, f(x) \geq t\} \quad \text{almost-everywhere,}$$

*and the sets $\{x, f(x) > t\}$ and $\{x, f(x) \geq t\}$ are both solutions of (5.3). In addition, under assumption **A**₂, the solution is unique:*

$$\Omega_t^* = \{x, f(x) > t\} = \{x, f(x) \geq t\}.$$

Observe that the curve EM^* is always well-defined, since $\int_{f \geq t} (f(x) - t) dx = \int_{f > t} (f(x) - t) dx$. We also point out that $EM^*(t) = \alpha(t) - t\lambda(t)$ for all $t > 0$, where we set $\alpha = \alpha_f$ and $\lambda = \lambda_f$ where α_f and λ_f are defined in (5.1).

Proposition 5.3. (DERIVATIVE AND CONVEXITY OF EM^*) Suppose that assumptions \mathbf{A}_1 and \mathbf{A}_2 are fulfilled. Then, the mapping EM^* is differentiable and we have for all $t > 0$:

$$EM^{*'}(t) = -\lambda(t).$$

In addition, the mapping $t > 0 \mapsto \lambda(t)$ being decreasing, the curve EM^* is convex.

We now introduce the concept of Excess-Mass curve of a scoring function $s \in \mathcal{S}$.

Definition 5.4. (*EM CURVES*) The EM curve of $s \in \mathcal{S}$ w.r.t. the probability distribution $F(dx)$ of a random variable X is the plot of the mapping

$$EM_s : t \in [0, \infty[\mapsto \sup_{A \in \{(\Omega_{s,t})_{t>0}\}} \mathbb{P}(X \in A) - t \text{Leb}(A), \quad (5.4)$$

where $\Omega_{s,t} = \{x \in \mathbb{R}^d, s(x) \geq t\}$ for all $t > 0$. One may also write: $\forall t > 0, EM_s(t) = \sup_{u>0} \alpha_s(u) - t\lambda_s(u)$. Finally, under assumption \mathbf{A}_1 , we have $EM_s(t) = 0$ for every $t > \|f\|_\infty$.

Regarding anomaly scoring, the concept of EM curve naturally induces a partial order on the set of all scoring functions: $\forall (s_1, s_2) \in \mathcal{S}^2$, s_1 is said to be more accurate than s_2 when $\forall t > 0, EM_{s_1}(t) \geq EM_{s_2}(t)$. Observe also that the optimal EM curve introduced in Definition 5.1 is itself the EM curve of a scoring function, the EM curve of any strictly increasing transform of the density f namely: $EM^* = EM_f$. Hence, in the unsupervised framework, optimal scoring functions are those maximizing the EM curve everywhere. In addition, maximizing EM_s can be viewed as recovering a collection of subsets $(\Omega_t^*)_{t>0}$ with maximum mass when penalized by their volume in a linear fashion. An optimal scoring function is then any $s \in \mathcal{S}$ with the Ω_t^{*} 's as level sets, for instance any scoring function of the form

$$s(x) = \int_{t=0}^{+\infty} \mathbf{1}_{x \in \Omega_t^*} a(t) dt, \quad (5.5)$$

with $a(t) > 0$ (observe that $s(x) = f(x)$ for $a \equiv 1$).

Proposition 5.5. (NATURE OF ANOMALY SCORING) Let $s \in \mathcal{S}$. The following properties hold true.

- (i) The mapping EM_s is non increasing on $(0, +\infty)$, takes its values in $[0, 1]$ and satisfies, $EM_s(t) \leq EM^*(t)$ for all $t \geq 0$.
- (ii) For $t \geq 0$, we have:

$$\inf_{u>0} \epsilon \text{Leb}(\{s > u\} \Delta_\epsilon \{f > t\}) \leq EM^*(t) - EM_s(t) \leq \|f\|_\infty \inf_{u>0} \text{Leb}(\{s > u\} \Delta \{f > t\}),$$

where $\{s > u\} \Delta_\epsilon \{f > t\} := \{f > t + \epsilon\} \setminus \{s > u\} \sqcup \{s > u\} \setminus \{f > t - \epsilon\}$ should be interpreted as a symmetric difference with ‘an ϵ tolerance’.

(iii) Let $\epsilon > 0$. Suppose that the quantity $\sup_{u>\epsilon} \int_{f^{-1}(\{u\})} 1/\|\nabla f(x)\| d\mu(x)$ is bounded, where μ denotes the $(d-1)$ -dimensional Hausdorff measure. Set $\epsilon_1 := \inf_T \|f - T \circ s\|_\infty$, where the infimum is taken over the set \mathcal{T} of all borelian increasing transforms $T : \mathbb{R}_+ \rightarrow \mathbb{R}_+$. Then,

$$\sup_{t \in [\epsilon + \epsilon_1, \|f\|_\infty]} |EM^*(t) - EM_s(t)| \leq C_1 \inf_{T \in \mathcal{T}} \|f - T \circ s\|_\infty,$$

where $C_1 = C(\epsilon_1, f)$ is a constant independent from $s(x)$.

Assertion (ii) provides a control of the pointwise difference between the optimal EM curve and EM_s in terms of the error made when recovering a specific minimum volume set Ω_t^* by a level set of $s(x)$. Thus the quantity $EM^*(t) - EM_s(t)$ measures how well level sets of s can approximate those of the underlying density. Assertion (iii) reveals that, if a certain increasing transform of a given scoring function $s(x)$ approximates well the density $f(x)$, then $s(x)$ is an accurate scoring function *w.r.t.* the EM criterion. As the distribution $F(dx)$ is generally unknown, EM curves must be estimated. Let $s \in \mathcal{S}$ and X_1, \dots, X_n be an i.i.d. sample with common distribution $F(dx)$ and set $\hat{\alpha}_s(t) = (1/n) \sum_{i=1}^n \mathbf{1}_{s(X_i) \geq t}$. The empirical EM curve of s is then defined as

$$\widehat{EM}_s(t) = \sup_{u>0} \{\hat{\alpha}_s(u) - t\lambda_s(u)\}.$$

In practice, it may be difficult to estimate the volume $\lambda_s(u)$ and Monte-Carlo approximation can naturally be used for this purpose.

5.4 A general approach to learn a scoring function

The concept of EM -curve provides a simple way to compare scoring functions but optimizing such a functional criterion is far from straightforward. As in Cl  men  on & Jakubowicz (2013), we propose to discretize the continuum of optimization problems and to construct a nearly optimal scoring function with level sets built by solving a finite collection of empirical versions of problem (5.3) over a subclass \mathcal{G} of borelian subsets. In order to analyze the accuracy of this approach, we introduce the following additional assumptions.

A₃ All minimum volume sets belong to \mathcal{G} :

$$\forall t > 0, \Omega_t^* \in \mathcal{G}.$$

A₄ The Rademacher average

$$\mathcal{R}_n = \mathbb{E} \left[\sup_{\Omega \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n \epsilon_i \mathbf{1}_{X_i \in \Omega} \right| \right]$$

is of order $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$, where $(\epsilon_i)_{i \geq 1}$ is a Rademacher chaos independent of the X_i 's.

Assumption **A₄** is very general and is fulfilled in particular when \mathcal{G} is of finite VC dimension, see Koltchinskii (2006), whereas the zero bias assumption **A₃** is in contrast very restrictive. It will be relaxed in Section 5.5.

Let $\delta \in (0, 1)$ and consider the complexity penalty $\Phi_n(\delta) = 2\mathcal{R}_n + \sqrt{\frac{\log(1/\delta)}{2n}}$. We have for all $n \geq 1$:

$$\mathbb{P} \left(\left\{ \sup_{G \in \mathcal{G}} (|P(G) - P_n(G)| - \Phi_n(\delta)) > 0 \right\} \right) \leq \delta, \quad (5.6)$$

see Koltchinskii (2006) for instance. Denote by $F_n = (1/n) \sum_{i=1}^n \delta_{X_i}$ the empirical measure based on the training sample X_1, \dots, X_n . For $t \geq 0$, define also the signed measures:

$$H_t(\cdot) = F(\cdot) - t\text{Leb}(\cdot) \\ \text{and} \quad H_{n,t}(\cdot) = F_n(\cdot) - t\text{Leb}(\cdot).$$

Equipped with these notations, for any $s \in \mathcal{S}$, we point out that one may write $EM^*(t) = \sup_{u \geq 0} H_t(\{x \in \mathbb{R}^d, f(x) \geq u\})$ and $EM_s(t) = \sup_{u \geq 0} H_t(\{x \in \mathbb{R}^d, s(x) \geq u\})$. Let $K > 0$ and $0 < t_K < t_{K-1} < \dots < t_1$. For k in $\{1, \dots, K\}$, let $\hat{\Omega}_{t_k}$ be an *empirical t_k -cluster*, that is to say a borelian subset of \mathbb{R}^d such that

$$\hat{\Omega}_{t_k} \in \arg \max_{\Omega \in \mathcal{G}} H_{n,t_k}(\Omega).$$

The empirical excess mass at level t_k is then $H_{n,t_k}(\hat{\Omega}_{t_k})$. The following result reveals the benefit of viewing density level sets as solutions of (5.3) rather than solutions of (5.2) (corresponding to a different parametrization of the thresholds).

Proposition 5.6. (MONOTONICITY) *For any k in $\{1, \dots, K\}$, the subsets $\cup_{i \leq k} \hat{\Omega}_{t_i}$ and $\cap_{i \geq k} \hat{\Omega}_{t_i}$ are still empirical t_k -clusters, just like $\hat{\Omega}_{t_k}$:*

$$H_{n,t_k}(\cup_{i \leq k} \hat{\Omega}_{t_i}) = H_{n,t_k}(\cap_{i \geq k} \hat{\Omega}_{t_i}) = H_{n,t_k}(\hat{\Omega}_{t_k}).$$

The result above shows that monotonous (regarding the inclusion) collections of empirical clusters can always be built. Coming back to the example depicted by Figure 5.3, as t decreases, the $\hat{\Omega}_t$'s are successively equal to A_1 , $A_1 \cup A_3$, and $A_1 \cup A_3 \cup A_2$, and are thus monotone as expected. This way, one fully avoids the problem inherent to the prior specification of a subdivision of the mass-axis in the MV -curve minimization approach (see the discussion in Section 5.2).

Consider an increasing sequence of empirical t_k clusters $(\hat{\Omega}_{t_k})_{1 \leq k \leq K}$ and a scoring function $s \in \mathcal{S}$ of the form

$$s_K(x) := \sum_{k=1}^K a_k \mathbf{1}_{x \in \hat{\Omega}_{t_k}}, \quad (5.7)$$

where $a_k > 0$ for every $k \in \{1, \dots, K\}$. Notice that the scoring function (5.7) can be seen as a Riemann sum approximation of (5.5) when $a_k = a(t_k) - a(t_{k+1})$. For simplicity solely, we take $a_k = t_k - t_{k+1}$ so that the $\hat{\Omega}_{t_k}$'s are t_k -level sets of s_K , i.e. $\hat{\Omega}_{t_k} = \{s \geq t_k\}$ and $\{s \geq t\} = \hat{\Omega}_{t_k}$ if $t \in]t_{k+1}, t_k]$. Observe that the results established in this work remain true for other choices. In the asymptotic framework considered in the subsequent analysis, it is stipulated that $K = K_n \rightarrow \infty$ as $n \rightarrow +\infty$. We assume in addition that $\sum_{k=1}^{\infty} a_k < \infty$.

Remark 5.7. (NESTED SEQUENCES) For $L \leq K$, we have $\{\Omega_{s_L, l}, l \geq 0\} = (\hat{\Omega}_{t_k})_{0 \leq k \leq L} \subset (\hat{\Omega}_{t_k})_{0 \leq k \leq K} = \{\Omega_{s_K, l}, l \geq 0\}$, so that by definition, $EM_{s_L} \leq EM_{s_K}$.

Remark 5.8. (RELATED WORK) We point out that a very similar result is proved in Polonik (1998) (see Lemma 2.2 therein) concerning the Lebesgue measure of the symmetric differences of density clusters.

Remark 5.9. (ALTERNATIVE CONSTRUCTION) It is noteworthy that, in practice, one may solve the optimization problems $\tilde{\Omega}_{t_k} \in \arg \max_{\Omega \in \mathcal{G}} H_{n, t_k}(\Omega)$ and next form $\hat{\Omega}_{t_k} = \cup_{i \leq k} \tilde{\Omega}_{t_i}$.

The following theorem provides rate bounds describing the performance of the scoring function s_K thus built with respect to the EM curve criterion in the case where the density f has compact support.

Theorem 5.10. (COMPACT SUPPORT CASE) Assume that conditions **A₁**, **A₂**, **A₃** and **A₄** hold true, and that f has a compact support. Let $\delta \in]0, 1[$, let $(t_k)_{k \in \{1, \dots, K\}}$ be such that $\sup_{1 \leq k < K} (t_k - t_{k+1}) = \mathcal{O}(1/\sqrt{n})$. Then, there exists a constant A independent from the t_k 's, n and δ such that, with probability at least $1 - \delta$, we have:

$$\sup_{t \in]0, t_1]} |EM^*(t) - EM_{s_K}(t)| \leq \left(A + \sqrt{2 \log(1/\delta)} + \text{Leb}(\text{supp} f) \right) \frac{1}{\sqrt{n}}.$$

Remark 5.11. (LOCALIZATION) The problem tackled in this work is that of scoring anomalies, which correspond to observations lying outside of ‘large’ excess mass sets, namely density clusters with parameter t close to zero. It is thus essential to establish rate bounds for the quantity $\sup_{t \in]0, C[} |EM^*(t) - EM_{s_K}(t)|$, where $C > 0$ depends on the proportion of the ‘least normal’ data we want to score/rank.

Proof of Theorem 5.10 (Sketch of). The proof results from the following lemma, which does not use the compact support assumption on f and is the starting point of the extension to the non compact support case (see Section 5.5.1).

Lemma 5.12. Suppose that assumptions **A₁**, **A₂**, **A₃** and **A₄** are fulfilled. Then, for $1 \leq k \leq K - 1$, there exists a constant A independent from n and δ , such that, with probability at least $1 - \delta$, for t in $]t_{k+1}, t_k]$,

$$|EM^*(t) - EM_{s_K}(t)| \leq \left(A + \sqrt{2 \log(1/\delta)} \right) \frac{1}{\sqrt{n}} + \lambda(t_{k+1})(t_k - t_{k+1}).$$

The detailed proof of this lemma is in the Detailed Proofs Section 5.9, and is a combination on the two following results, the second one being a straightforward consequence of the derivative property of EM^* (Proposition 5.3):

- With probability at least $1 - \delta$, for $k \in \{1, \dots, K\}$,

$$0 \leq EM^*(t_k) - EM_{s_K}(t_k) \leq 2\Phi_n(\delta) .$$

- Let k in $\{1, \dots, K - 1\}$. Then for every t in $]t_{k+1}, t_k]$,

$$0 \leq EM^*(t) - EM^*(t_k) \leq \lambda(t_{k+1})(t_k - t_{k+1}) .$$

□

5.5 Extensions - Further results

This section is devoted to extend the results of the previous one. We first relax the compact support assumption and next the one stipulating that all density level sets belong to the class \mathcal{G} , namely \mathbf{A}_3 .

5.5.1 Distributions with non compact support

It is the purpose of this section to show that the algorithm detailed below produces a scoring function s such that EM_s is uniformly close to EM^* (Theorem 5.13). See Figure 5.3 as an illustration and a comparison with the MV formulation as used as a way to recover empirical minimum volume set $\hat{\Gamma}_\alpha$.

The main argument to extend the above results to the case where $\text{supp} f$ is not bounded is given in Lemma 5.12 in Section 5.9. The meshgrid (t_k) must be chosen adaptively, in a data-driven fashion. Let $h : \mathbb{R}_+^* \rightarrow \mathbb{R}_+$ be a decreasing function such that $\lim_{t \rightarrow 0} h(t) = +\infty$. Just like the previous approach, the grid is described by a decreasing sequence (t_k) . Let $t_1 \geq 0$, $N > 0$ and define recursively $t_1 > t_2 > \dots > t_N > t_{N+1} = 0$, as well as $\hat{\Omega}_{t_1}, \dots, \hat{\Omega}_{t_N}$, through

$$t_{k+1} = t_k - (\sqrt{n})^{-1} \frac{1}{h(t_{k+1})} \tag{5.9}$$

$$\hat{\Omega}_{t_k} = \arg \max_{\Omega \in \mathcal{G}} H_{n,t_k}(\Omega), \tag{5.10}$$

Algorithm 2 Learning a scoring function

Suppose that assumptions $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4$ hold true.

Let t_1 such that $\max_{\Omega \in \mathcal{G}} H_{n,t_1}(\Omega) \geq 0$. Fix $N > 0$. For $k = 1, \dots, N$,

1. Find $\tilde{\Omega}_{t_k} \in \arg \max_{\Omega \in \mathcal{G}} H_{n,t_k}(\Omega)$,
2. Define $\hat{\Omega}_{t_k} = \cup_{i \leq k} \tilde{\Omega}_{t_i}$
3. Set $t_{k+1} = \frac{t_1}{(1 + \frac{1}{\sqrt{n}})^k}$ for $k \leq N - 1$.

In order to reduce the complexity, we may replace steps 1 and 2 with

$$\hat{\Omega}_{t_k} \in \arg \max_{\Omega \supset \hat{\Omega}_{t_{k-1}}} H_{n,t_k}(\Omega).$$

The resulting piecewise constant scoring function is

$$s_N(x) = \sum_{k=1}^N (t_k - t_{k+1}) \mathbb{1}_{x \in \hat{\Omega}_{t_k}}. \quad (5.8)$$

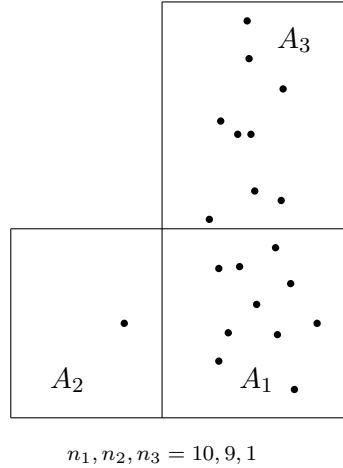


FIGURE 5.3: Unsuccessful mass-volume criterion optimization

Sample of $n = 20$ points in a 2-d space, partitioned into three rectangles. As α increases, the minimum volume sets $\hat{\Gamma}_\alpha$ are successively equal to A_1 , $A_1 \cup A_2$, $A_1 \cup A_3$, and $A_1 \cup A_3 \cup A_2$, whereas, in the excess-mass approach, as t decreases, the $\hat{\Omega}_t$'s are successively equal to A_1 , $A_1 \cup A_3$, and $A_1 \cup A_3 \cup A_2$.

with the property that $\hat{\Omega}_{t_{k+1}} \supset \hat{\Omega}_{t_k}$. As pointed out in Remark 5.9, it suffices to take $\hat{\Omega}_{t_{k+1}} = \tilde{\Omega}_{t_{k+1}} \cup \hat{\Omega}_{t_k}$, where $\tilde{\Omega}_{t_{k+1}} = \arg \max_{\Omega \in \mathcal{G}} H_{n,t_k}(\Omega)$. This yields the scoring function s_N defined by (5.8) such that by virtue of Lemma 5.12 (see technical details in Section 5.9), with probability at least $1 - \delta$,

$$\sup_{t \in [t_N, t_1]} |EM^*(t) - EM_{s_N}(t)| \leq \left(A + \sqrt{2 \log(1/\delta)} + \sup_{1 \leq k \leq N} \frac{\lambda(t_k)}{h(t_k)} \right) \frac{1}{\sqrt{n}}.$$

Therefore, if we take h such that $\lambda(t) = \mathcal{O}(h(t))$ as $t \rightarrow 0$, we can assume that $\lambda(t)/h(t) \leq B$ for t in $]0, t_1]$ since λ is decreasing, and we obtain:

$$\sup_{t \in]t_N, t_1]} |EM^*(t) - EM_{s_N}(t)| \leq \left(A + \sqrt{2 \log(1/\delta)} \right) \frac{1}{\sqrt{n}}. \quad (5.11)$$

On the other hand from $t \text{Leb}(\{f > t\}) \leq \int_{f>t} f \leq 1$, we have $\lambda(t) \leq 1/t$. Thus h can be chosen as $h(t) := 1/t$ for $t \in]0, t_1]$. In this case, (5.10) yields, for $k \geq 2$,

$$t_k = \frac{t_1}{\left(1 + \frac{1}{\sqrt{n}}\right)^{k-1}}. \quad (5.12)$$

Theorem 5.13. (UNBOUNDED SUPPORT CASE) *Suppose that assumptions \mathbf{A}_1 , \mathbf{A}_2 , \mathbf{A}_3 , \mathbf{A}_4 hold true, let $t_1 > 0$ and for $k \geq 2$, consider t_k as defined by (5.12), Ω_{t_k} by (5.9), and s_N (5.8). Then there is a constant A independent from N , n and δ such that, with probability larger than $1 - \delta$, we have:*

$$\sup_{t \in]0, t_1]} |EM^*(t) - EM_{s_N}(t)| \leq \left[A + \sqrt{2 \log(1/\delta)} \right] \frac{1}{\sqrt{n}} + o_N(1),$$

where $o_N(1) = 1 - EM^*(t_N)$. In addition, $s_N(x)$ converges to $s_\infty(x) := \sum_{k=1}^{\infty} (t_{k+1} - t_k) \mathbb{1}_{\Omega_{t_{k+1}}}$ as $N \rightarrow \infty$ and s_∞ is such that, for all $\delta \in (0, 1)$, we have with probability at least $1 - \delta$:

$$\sup_{t \in]0, t_1]} |EM^*(t) - EM_{s_\infty}(t)| \leq \left[A + \sqrt{2 \log(1/\delta)} \right] \frac{1}{\sqrt{n}}$$

Proof of Theorem 5.13 (Sketch of). The first assertion is a consequence of (5.11) combined with the fact that

$$\begin{aligned} \sup_{t \in]0, t_N]} |EM^*(t) - EM_{s_N}(t)| &\leq 1 - EM_{s_N}(t_N) \\ &\leq 1 - EM^*(t_N) + 2\Phi_n(\delta) \end{aligned}$$

holds true with probability at least $1 - \delta$. For the second part, it suffices to observe that $s_N(x)$ (absolutely) converges to s_∞ and that, as pointed out in Remark 5.7, $EM_{s_N} \leq EM_{s_\infty}$. For a detailed proof, see Section 5.9. \square

5.5.2 Bias analysis

In this subsection, we relax assumption \mathbf{A}_3 . For any collection \mathcal{C} of subsets of \mathbb{R}^d , $\sigma(\mathcal{C})$ denotes here the σ -algebra generated by \mathcal{C} . Consider the hypothesis below.

$\tilde{\mathbf{A}}_3$ *There exists a countable subcollection of \mathcal{G} , $F = \{F_i\}_{i \geq 1}$ say, forming a partition of \mathbb{R}^d and such that $\sigma(F) \subset \mathcal{G}$.*

Denote by f_F the best approximation (for the L_2 -norm) of f by piecewise functions on F ,

$$f_F(x) := \sum_{i \geq 1} \mathbb{1}_{x \in F_i} \frac{1}{\text{Leb}(F_i)} \int_{F_i} f(y) dy .$$

Then, variants of Theorems 5.10 and 5.13 can be established without assumption \mathbf{A}_3 , as soon as $\tilde{\mathbf{A}}_3$ holds true, at the price of the additional term $\|f - f_F\|_{L^1}$ in the bound, related to the inherent bias. For illustration purpose, the following result generalizes one of the inequalities stated in Theorem 5.13:

Theorem 5.14. (BIASED EMPIRICAL CLUSTERS) *Suppose that assumptions \mathbf{A}_1 , \mathbf{A}_2 , $\tilde{\mathbf{A}}_3$, \mathbf{A}_4 hold true, let $t_1 > 0$ and for $k \geq 2$ consider t_k defined by (5.12), Ω_{t_k} by (5.9), and s_N by (5.8). Then there is a constant A independent from N , n , δ such that, with probability larger than $1 - \delta$, we have:*

$$\sup_{t \in]0, t_1]} |EM^*(t) - EM_{s_N}(t)| \leq \left[A + \sqrt{2 \log(1/\delta)} \right] \frac{1}{\sqrt{n}} + \|f - f_F\|_{L^1} + o_N(1),$$

where $o_N(1) = 1 - EM^*(t_N)$.

Remark 5.15. (HYPERCUBES) In practice, one defines a sequence of models $F_l \subset \mathcal{G}_l$ indexed by a tuning parameter l controlling (the inverse of) model complexity, such that $\|f - f_{F_l}\|_{L^1} \rightarrow 0$ as $l \rightarrow 0$. For instance, the class F_l could be formed by disjoint hypercubes of side length l .

Proof of Theorem 5.14 (Sketch of). The result directly follows from the following lemma, which establishes an upper bound for the bias, with the notations $EM_{\mathcal{C}}^*(t) := \max_{\Omega \in \mathcal{C}} H_t(\Omega) \leq EM^*(t) = \max_{\Omega \text{ meas.}} H_t(\Omega)$ for any class of measurable sets \mathcal{C} , and $\mathcal{F} := \sigma(F)$ so that by assumption \mathbf{A}_3 , $\mathcal{F} \subset \mathcal{G}$. Details are omitted due to space limits.

Lemma 5.16. *Under assumption $\tilde{\mathbf{A}}_3$, we have for every t in $[0, \|f\|_{\infty}]$,*

$$0 \leq EM^*(t) - EM_{\mathcal{F}}^*(t) \leq \|f - f_F\|_{L^1} .$$

The model bias $EM^ - EM_{\mathcal{G}}^*$ is then uniformly bounded by $\|f - f_F\|_{L^1}$.*

To prove this lemma (see Section 5.9 for details), one shows that:

$$EM^*(t) - EM_{\mathcal{F}}^*(t) \leq \int_{f > t} (f - f_F) + \int_{\{f > t\} \setminus \{f_F > t\}} (f_F - t) - \int_{\{f_F > t\} \setminus \{f > t\}} (f_F - t) ,$$

where we use the fact that for all $t > 0$, $\{f_F > t\} \in \mathcal{F}$ and $\forall F \in \mathcal{F}$, $\int_G f = \int_G f_F$. It suffices then to observe that the second and the third term in the bound are non-positive. \square

5.6 Simulation examples

Algorithm 2 is here implemented from simulated 2-*d heavy-tailed* data with common density $f(x, y) = 1/2 \times 1/(1 + |x|)^3 \times 1/(1 + |y|)^2$. The training set is of size $n = 10^5$, whereas the test set counts 10^6 points. For $l > 0$, we set $\mathcal{G}_l = \sigma(F)$ where $F_l = \{F_i^l\}_{i \in \mathbb{Z}^2}$ and $F_i^l = [li_1, li_1 + 1] \times [li_2, li_2 + 1]$ for all $i = (i_1, i_2) \in \mathbb{Z}^2$. The bias of the model is thus bounded by $\|f - f_F\|_\infty$, vanishing as $l \rightarrow 0$ (observe that the bias is at most of order l as soon as f is Lipschitz for instance). The scoring function s is built using the points located in $[-L, L]^2$ and setting $s = 0$ outside of $[-L, L]^2$. Practically, one takes L as the maximum norm value of the points in the training set, or such that an empirical estimate of $\mathbb{P}(X \in [-L, L]^2)$ is very close to 1 (here one obtains 0.998 for $L = 500$). The implementation of our algorithm involves the use of a sparse matrix to store the data in the partition of hypercubes, such that the complexity of the procedure for building the scoring function s and that of the computation of its empirical EM -curve is very small compared to that needed to compute f_{F_l} and $EM_{f_{F_l}}$, which are given here for the sole purpose of quantifying the model bias.

Figure 5.4 illustrates as expected the deterioration of EM_s for large l , except for t close to zero: this corresponds to the model bias. However, Figure 5.5 reveals an ‘overfitting’ phenomenon for values of t close to zero, when l is fairly small. This is mainly due to the fact that subsets involved in the scoring function are then tiny in regions where there are very few observations (in the tail of the distribution). On the other hand, for the largest values of t , the smallest values of l give the best results: the smaller the parameter l , the weaker the model bias and no overfitting is experienced because of the high local density of the observations. Recalling the notation $EM_{\mathcal{G}}^*(t) = \max_{\Omega \in \mathcal{G}} H_t(\Omega) \leq EM^*(t) = \max_{\Omega \text{ meas.}} H_t(\Omega)$ so that the bias of our model is $EM^* - EM_{\mathcal{G}}^*$, Figure 5.6 illustrates the variations of the bias with the wealth of our model characterized by l the width of the partition by hypercubes. Notice that partitions with small l are not so good approximation for large t , but are performing as well as the other in the extreme values, namely when t is close to 0. On the top of that, those partitions have the merit not to overfit the extreme datas, which typically are isolated.

This empirical analysis demonstrates that introducing a notion of adaptivity for the partition F , with progressively growing bin-width as t decays to zero and as the hypercubes are being selected in the construction of s (which crucially depends on local properties of the empirical distribution), drastically improves the accuracy of the resulting scoring function in the EM curve sense.

5.7 Conclusion

Prolongating the contribution of Cl  men  on & Jakubowicz (2013), this chapter provides an alternative view (respectively, an other parameterization) of the anomaly scoring problem, leading

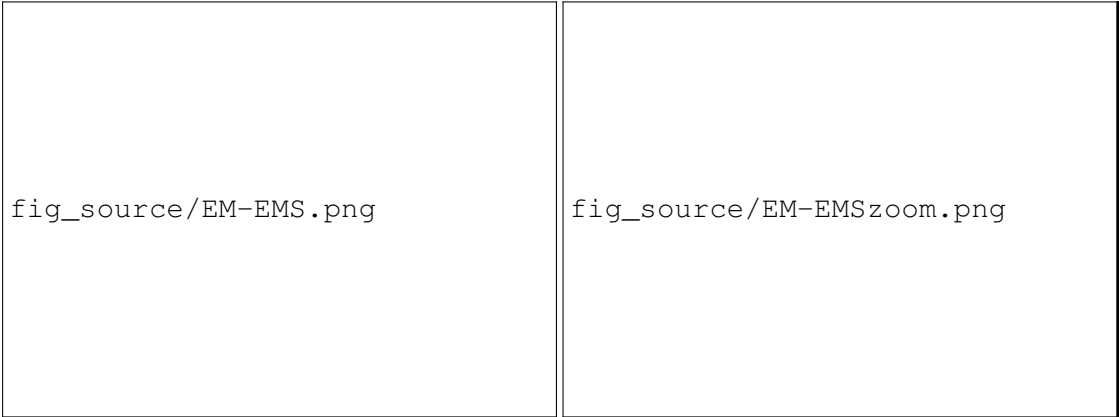


FIGURE 5.4: Optimal and realized EM curves

FIGURE 5.5: Zoom near 0

FIGURE 5.6: EM_G for different l

to another adaptive method to build scoring functions, which offers theoretical and computational advantages both at the same time. This novel formulation yields a procedure producing a nested sequence of empirical density level sets, and exhibits a good performance, even in the non compact support case. Thus, the main drawbacks of the mass-volume curve criterion listed in the introduction section are resolved excepting drawback **1**). In addition, the model bias has been incorporated in the rate bound analysis. However, the use of the Excess-Mass criterion to measure the quality of a scoring function s_n involves the computation of the Lebesgue measure $\text{Leb}(s_n \geq u)$, just as with the Mass-Volume criterion (drawback **1**). This is a major drawback

for its use in high dimensional framework, if no prior knowledge on the form of these level sets is available.

5.8 Illustrations

Note that the scoring function we built in Algorithm 2 is incidentally an estimator of the density f (usually called the silhouette), since $f(x) = \int_0^\infty \mathbb{1}_{f \geq t} dt = \int_0^\infty \mathbb{1}_{\Omega_t^*} dt$ and $s(x) := \sum_{k=1}^K (t_k - t_{k-1}) \mathbb{1}_{x \in \hat{\Omega}_{t_k}}$ which is a discretization of $\int_0^\infty \mathbb{1}_{\hat{\Omega}_t} dt$. This fact is illustrated in Figure 5.7. Note that the silhouette does not focus on local properties of the density, but only on its induced pre-order (level sets).



FIGURE 5.7: density and scoring functions

5.9 Detailed Proofs

Proof of Proposition 5.3

Let $t > 0$. Recall that $EM^*(t) = \alpha(t) - t\lambda(t)$ where $\alpha(t)$ denote the mass at level t , namely $\alpha(t) = \mathbb{P}(f(X) \geq t)$, and $\lambda(t)$ denote the volume at level t , i.e. $\lambda(t) = \text{Leb}(\{x, f(x) \geq t\})$. For $h > 0$, let $A(h)$ denote the quantity

$$A(h) = \frac{1}{h}(\alpha(t+h) - \alpha(t))$$

and

$$B(h) = \frac{1}{h}(\lambda(t+h) - \lambda(t)).$$

It is straightforward to see that $A(h)$ and $B(h)$ converge when $h \rightarrow 0$, and expressing $EM^{*'} = \alpha'(t) - t\lambda'(t) - \lambda(t)$, it suffices to show that $\alpha'(t) - t\lambda'(t) = 0$, namely $\lim_{h \rightarrow 0} A(h) - t B(h) = 0$. Yet, we have

$$A(h) - t B(h) = \frac{1}{h} \int_{t \leq f \leq t+h} f - t \leq \frac{1}{h} \int_{t \leq f \leq t+h} h = \text{Leb}(t \leq f \leq t+h) \rightarrow 0$$

because f has no flat part.

Proof of Lemma 5.2:

On the one hand, for every Ω measurable,

$$\begin{aligned} \mathbb{P}(X \in \Omega) - t \text{Leb}(\Omega) &= \int_{\Omega} (f(x) - t) dx \\ &\leq \int_{\Omega \cap \{f \geq t\}} (f(x) - t) dx \\ &\leq \int_{\{f \geq t\}} (f(x) - t) dx \\ &= \mathbb{P}(f(X) \geq t) - t \text{Leb}(\{f \geq t\}). \end{aligned}$$

It follows that $\{f \geq t\} \in \arg \max_{A \text{ meas.}} \mathbb{P}(X \in A) - t \text{Leb}(A)$.

On the other hand, suppose $\Omega \in \arg \max_{A \text{ meas.}} \mathbb{P}(X \in A) - t \text{Leb}(A)$ and $\text{Leb}(\{f > t\} \setminus \Omega) > 0$. Then there is $\epsilon > 0$ such that $\text{Leb}(\{f > t + \epsilon\} \setminus \Omega) > 0$ (by sub-additivity of Leb , if it is not the case, then $\text{Leb}(\{f > t\} \setminus \Omega) = \text{Leb}(\cup_{\epsilon \in \mathbb{Q}_+} \{f > t + \epsilon\} \setminus \Omega) = 0$). We have thus

$$\int_{\{f > t\} \setminus \Omega} (f(x) - t) dx > \epsilon \cdot \text{Leb}(\{f > t + \epsilon\} \setminus \Omega) > 0,$$

so that

$$\begin{aligned} \int_{\Omega} (f(x) - t) dx &\leq \int_{\{f > t\}} (f(x) - t) dx - \int_{\{f > t\} \setminus \Omega} (f(x) - t) dx \\ &< \int_{\{f > t\}} (f(x) - t) dx, \end{aligned}$$

i.e

$$\mathbb{P}(X \in \Omega) - t \text{Leb}(\Omega) < \mathbb{P}(f(X) \geq t) - t \text{Leb}(\{x, f(x) \geq t\})$$

which is a contradiction. Thus, $\{f > t\} \subset \Omega$ Leb -almost surely.

To show that $\Omega_t^* \subset \{x, f(x) \geq t\}$, suppose that $\text{Leb}(\Omega_t^* \cap \{f < t\}) > 0$. Then by sub-additivity of Leb just as above, there is $\epsilon > 0$ such that $\text{Leb}(\Omega_t^* \cap \{f < t - \epsilon\}) > 0$ and

$$\int_{\Omega_t^* \cap \{f < t - \epsilon\}} f - t \leq -\epsilon \cdot \text{Leb}(\Omega_t^* \cap \{f < t - \epsilon\}) < 0.$$

It follows that

$$\mathbb{P}(X \in \Omega_t^*) - t \text{Leb}(\Omega_t^*) < \mathbb{P}(X \in \Omega_t^* \setminus \{f < t - \epsilon\}) - t \text{Leb}(\Omega_t^* \setminus \{f < t - \epsilon\}),$$

which is a contradiction with the optimality of Ω_t^* .

Proof of Proposition 5.5

Proving the first assertion is immediate, since $\int_{f \geq t} (f(x) - t)dx \geq \int_{s \geq t} (f(x) - t)dx$. Let us now turn to the second assertion. We have:

$$\begin{aligned} EM^*(t) - EM_s(t) &= \int_{f > t} (f(x) - t)dx - \sup_{u > 0} \int_{s > u} (f(x) - t)dx \\ &= \inf_{u > 0} \int_{f > t} (f(x) - t)dx - \int_{s > u} (f(x) - t)dx. \end{aligned}$$

Yet,

$$\begin{aligned} &\int_{\{f > t\} \setminus \{s > u\}} (f(x) - t)dx + \int_{\{s > u\} \setminus \{f > t\}} (t - f(x))dx \\ &\leq (\|f\|_\infty - t) \cdot \text{Leb}(\{f > t\} \setminus \{s > u\}) + t \text{Leb}(\{s > u\} \setminus \{f > t\}), \end{aligned}$$

so we obtain:

$$\begin{aligned} EM^*(t) - EM_s(t) &\leq \max(t, \|f\|_\infty - t) \text{Leb}(\{s > u\} \Delta \{f > t\}) \\ &\leq \|f\|_\infty \cdot \text{Leb}(\{s > u\} \Delta \{f > t\}). \end{aligned}$$

To prove the third point, note that:

$$\inf_{u > 0} \text{Leb}(\{s > u\} \Delta \{f > t\}) = \inf_{T \nearrow} \text{Leb}(\{Ts > t\} \Delta \{f > t\})$$

Yet,

$$\begin{aligned} \text{Leb}\left(\{Ts > t\} \Delta \{f > t\}\right) &\leq \text{Leb}(\{f > t - \|Ts - f\|_\infty\} \setminus \{f > t + \|Ts - f\|_\infty\}) \\ &= \lambda(t - \|Ts - f\|_\infty) - \lambda(t + \|Ts - f\|_\infty) \\ &= - \int_{t - \|Ts - f\|_\infty}^{t + \|Ts - f\|_\infty} \lambda'(u) du. \end{aligned}$$

On the other hand, we have $\lambda(t) = \int_{\mathbb{R}^d} \mathbb{1}_{f(x) \geq t} dx = \int_{\mathbb{R}^d} g(x) \|\nabla f(x)\| dx$, where we let

$$g(x) = \frac{1}{\|\nabla f(x)\|} \mathbb{1}_{\{x, \|\nabla f(x)\| > 0, f(x) \geq t\}}.$$

The co-area formula (see Federer (1969), p.249, th.3.2.12) gives in this case:

$$\lambda(t) = \int_{\mathbb{R}} du \int_{f^{-1}(u)} \frac{1}{\|\nabla f(x)\|} \mathbb{1}_{\{x, f(x) \geq t\}} d\mu(x) = \int_t^\infty du \int_{f^{-1}(u)} \frac{1}{\|\nabla f(x)\|} d\mu(x)$$

so that $\lambda'(t) = - \int_{f^{-1}(t)} \frac{1}{\|\nabla f(x)\|} d\mu(x)$.

Let η_ϵ such that $\forall u > \epsilon$, $|\lambda'(u)| = \int_{f^{-1}(u)} \frac{1}{\|\nabla f(x)\|} d\mu(x) < \eta_\epsilon$. We obtain:

$$\sup_{t \in [\epsilon + \inf_{T \nearrow} \|f - Ts\|_\infty, \|f\|_\infty]} EM^*(t) - EM_s(t) \leq 2\eta_\epsilon \cdot \|f\|_\infty \inf_{T \nearrow} \|f - Ts\|_\infty.$$

In particular, if $\inf_{T \nearrow} \|f - Ts\|_\infty \leq \epsilon_1$,

$$\sup_{[\epsilon + \epsilon_1, \|f\|_\infty]} |EM^* - EM_s| \leq 2\eta_\epsilon \cdot \|f\|_\infty \cdot \inf_{T \nearrow} \|f - Ts\|_\infty.$$

Proof of Proposition 5.6

Let i in $\{1, \dots, K\}$. First, note that:

$$\begin{aligned} H_{n, t_{i+1}}(\hat{\Omega}_{t_{i+1}} \cup \hat{\Omega}_{t_i}) &= H_{n, t_{i+1}}(\hat{\Omega}_{t_{i+1}}) + H_{n, t_{i+1}}(\hat{\Omega}_{t_i} \setminus \hat{\Omega}_{t_{i+1}}), \\ H_{n, t_i}(\hat{\Omega}_{t_{i+1}} \cap \hat{\Omega}_{t_i}) &= H_{n, t_i}(\hat{\Omega}_{t_i}) - H_{n, t_i}(\hat{\Omega}_{t_i} \setminus \hat{\Omega}_{t_{i+1}}). \end{aligned}$$

It follows that

$$\begin{aligned} &H_{n, t_{i+1}}(\hat{\Omega}_{t_{i+1}} \cup \hat{\Omega}_{t_i}) + H_{n, t_i}(\hat{\Omega}_{t_{i+1}} \cap \hat{\Omega}_{t_i}) \\ &= H_{n, t_{i+1}}(\hat{\Omega}_{t_{i+1}}) + H_{n, t_i}(\hat{\Omega}_{t_i}) + H_{n, t_{i+1}}(\hat{\Omega}_{t_i} \setminus \hat{\Omega}_{t_{i+1}}) - H_{n, t_i}(\hat{\Omega}_{t_i} \setminus \hat{\Omega}_{t_{i+1}}), \end{aligned}$$

with $H_{n,t_{i+1}}(\hat{\Omega}_{t_i} \setminus \hat{\Omega}_{t_{i+1}}) - H_{n,t_i}(\hat{\Omega}_{t_i} \setminus \hat{\Omega}_{t_{i+1}}) \geq 0$ since $H_{n,t}$ is decreasing in t . But on the other hand, by definition of $\hat{\Omega}_{t_{i+1}}$ and $\hat{\Omega}_{t_i}$ we have:

$$\begin{aligned} H_{n,t_{i+1}}(\hat{\Omega}_{t_{i+1}} \cup \hat{\Omega}_{t_i}) &\leq H_{n,t_{i+1}}(\hat{\Omega}_{t_{i+1}}), \\ H_{n,t_i}(\hat{\Omega}_{t_{i+1}} \cap \hat{\Omega}_{t_i}) &\leq H_{n,t_i}(\hat{\Omega}_{t_i}). \end{aligned}$$

Finally we get:

$$\begin{aligned} H_{n,t_{i+1}}(\hat{\Omega}_{t_{i+1}} \cup \hat{\Omega}_{t_i}) &= H_{n,t_{i+1}}(\hat{\Omega}_{t_{i+1}}), \\ H_{n,t_i}(\hat{\Omega}_{t_{i+1}} \cap \hat{\Omega}_{t_i}) &= H_{n,t_i}(\hat{\Omega}_{t_i}). \end{aligned}$$

Proceeding by induction we have, for every m such that $k + m \leq K$:

$$\begin{aligned} H_{n,t_{i+m}}(\hat{\Omega}_{t_i} \cup \hat{\Omega}_{t_{i+1}} \cup \dots \cup \hat{\Omega}_{t_{i+m}}) &= H_{n,t_{i+m}}(\hat{\Omega}_{t_{i+m}}), \\ H_{n,t_i}(\hat{\Omega}_{t_i} \cap \hat{\Omega}_{t_{i+1}} \cap \dots \cap \hat{\Omega}_{t_{i+m}}) &= H_{n,t_i}(\hat{\Omega}_{t_i}). \end{aligned}$$

Taking $(i=1, m=k-1)$ for the first equation and $(i=k, m=K-k)$ for the second completes the proof.

Proof of Theorem 5.10

We shall use the following lemma:

Lemma 5.17. *With probability at least $1 - \delta$, for $k \in \{1, \dots, K\}$,*

$$0 \leq EM^*(t_k) - EM_{s_K}(t_k) \leq 2\Phi_n(\delta).$$

Proof of Lemma 5.17:

Remember that by definition of $\hat{\Omega}_{t_k}$: $H_{n,t_k}(\hat{\Omega}_{t_k}) = \max_{\Omega \in \mathcal{G}} H_{n,t_k}(\Omega)$ and note that:

$$EM^*(t_k) = \max_{\Omega \text{ meas.}} H_{t_k}(\Omega) = \max_{\Omega \in \mathcal{G}} H_{t_k}(\Omega) \geq H_{t_k}(\hat{\Omega}_{t_k}).$$

On the other hand, using (5.6), with probability at least $1 - \delta$, for every $G \in \mathcal{G}$, $|\mathbb{P}(G) - \mathbb{P}_n(G)| \leq \Phi_n(\delta)$. Hence, with probability at least $1 - \delta$, for all $\Omega \in \mathcal{G}$:

$$H_{n,t_k}(\Omega) - \Phi_n(\delta) \leq H_{t_k}(\Omega) \leq H_{n,t_k}(\Omega) + \Phi_n(\delta)$$

so that, with probability at least $(1 - \delta)$, for $k \in \{1, \dots, K\}$,

$$H_{n,t_k}(\hat{\Omega}_{t_k}) - \Phi_n(\delta) \leq H_{t_k}(\hat{\Omega}_{t_k}) \leq EM^*(t_k) \leq H_{n,t_k}(\hat{\Omega}_{t_k}) + \Phi_n(\delta),$$

whereby, with probability at least $(1 - \delta)$, for $k \in \{1, \dots, K\}$,

$$0 \leq EM^*(t_k) - H_{t_k}(\hat{\Omega}_{t_k}) \leq 2\Phi_n(\delta).$$

The following Lemma is a consequence of the derivative property of EM^* (Proposition 5.3)

Lemma 5.18. *Let k in $\{1, \dots, K - 1\}$. Then for every t in $]t_{k+1}, t_k]$,*

$$0 \leq EM^*(t) - EM^*(t_k) \leq \lambda(t_{k+1})(t_k - t_{k+1}).$$

Combined with Lemma 5.17 and the fact that EM_{s_K} is non-increasing, and writing

$$\begin{aligned} EM^*(t) - EM_{s_K}(t) &= (EM^*(t) - EM^*(t_k)) + (EM^*(t_k) - EM_{s_K}(t_k)) \\ &\quad + (EM_{s_K}(t_k) - EM_{s_K}(t)) \end{aligned}$$

this result leads to:

$$\forall k \in \{0, \dots, K - 1\}, \forall t \in]t_{k+1}, t_k],$$

$$0 \leq EM^*(t) - EM_{s_K}(t) \leq 2\Phi_n(\delta) + \lambda(t_{k+1})(t_k - t_{k+1})$$

which gives Lemma 5.12 stated in the sketch of proof. Notice that we have not yet used the fact that f has a compact support.

The compactness support assumption allows an extension of Lemma 5.18 to $k = K$, namely the inequality holds true for t in $]t_{K+1}, t_K] =]0, t_K]$ as soon as we let $\lambda(t_{K+1}) := \text{Leb}(\text{supp}f)$. Indeed the compactness of $\text{supp}f$ implies that $\lambda(t) \rightarrow \text{Leb}(\text{supp}f)$ as $t \rightarrow 0$. Observing that Lemma 5.17 already contains the case $k = K$, this leads to, for k in $\{0, \dots, K\}$ and $t \in]t_{k+1}, t_k]$, $|EM^*(t) - EM_{s_K}(t)| \leq 2\Phi_n(\delta) + \lambda(t_{k+1})(t_k - t_{k+1})$. Therefore, λ being a decreasing function bounded by $\lambda(\text{Leb}(\text{supp}f))$, we obtain the following: with probability at least $1 - \delta$, we have for all t in $]0, t_1]$,

$$|EM^*(t) - EM_{s_K}(t)| \leq \left(A + \sqrt{2\log(1/\delta)}\right) \frac{1}{\sqrt{n}} + \lambda(\text{Leb}(\text{supp}f)) \sup_{1 \leq k \leq K} (t_k - t_{k+1}).$$

Proof of Theorem 5.13

The first part of this theorem is a consequence of (5.11) combined with:

$$\sup_{t \in]0, t_N]} |EM^*(t) - EM_{s_N}(t)| \leq 1 - EM_{s_N}(t_N) \leq 1 - EM^*(t_N) + 2\Phi_n(\delta),$$

where we use the fact that $0 \leq EM^*(t_N) - EM_{s_N}(t_N) \leq 2\Phi_n(\delta)$ following from Lemma 5.17.

To see the convergence of $s_N(x)$, note that:

$$s_N(x) = \frac{t_1}{\sqrt{n}} \sum_{k=1}^{\infty} \frac{1}{(1 + \frac{1}{\sqrt{n}})^k} \mathbb{1}_{x \in \hat{\Omega}_{t_k}} \mathbb{1}_{\{k \leq N\}} \leq \frac{t_1}{\sqrt{n}} \sum_{k=1}^{\infty} \frac{1}{(1 + \frac{1}{\sqrt{n}})^k} < \infty,$$

and analogically to Remark 5.7 observe that $EM_{s_N} \leq EM_{s_\infty}$ so that $\sup_{t \in [0, t_1]} |EM^*(t) - EM_{s_\infty}(t)| \leq \sup_{t \in [0, t_1]} |EM^*(t) - EM_{s_N}(t)|$ which proves the last part of the theorem.

Proof of Lemma 5.16

By definition, for every class of set \mathcal{H} , $EM_{\mathcal{H}}^*(t) = \max_{\Omega \in \mathcal{H}} H_t(\Omega)$. The bias $EM^*(t) - EM_{\mathcal{G}}^*(t)$ of the model \mathcal{G} is majored by $EM^*(t) - EM_{\mathcal{F}}^*(t)$ since $\mathcal{F} \subset \mathcal{G}$. Remember that

$$f_F(x) := \sum_{i \geq 1} \mathbb{1}_{x \in F_i} \frac{1}{|F_i|} \int_{F_i} f(y) dy,$$

and note that for all $t > 0$, $\{f_F > t\} \in \mathcal{F}$. It follows that:

$$\begin{aligned} EM^*(t) - EM_{\mathcal{F}}^*(t) &= \int_{f > t} (f - t) - \sup_{C \in \mathcal{F}} \int_C (f - t) \\ &\leq \int_{f > t} (f - t) - \int_{f_F > t} (f - t) \quad \text{since } \{f_F > t\} \in \mathcal{F} \\ &= \int_{f > t} (f - t) - \int_{f_F > t} (f_F - t) \quad \text{since } \forall G \in \mathcal{F}, \int_G f = \int_G f_F \\ &= \int_{f > t} (f - t) - \int_{f > t} (f_F - t) + \int_{f > t} (f_F - t) - \int_{f_F > t} (f_F - t) \\ &= \int_{f > t} (f - f_F) + \int_{\{f > t\} \setminus \{f_F > t\}} (f_F - t) - \int_{\{f_F > t\} \setminus \{f > t\}} (f_F - t). \end{aligned}$$

Observe that the second and the third term in the bound are non-positive. Therefore:

$$EM^*(t) - EM_{\mathcal{F}}^*(t) \leq \int_{f > t} (f - f_F) \leq \int_{\mathbb{R}^d} |f - f_F|.$$

CHAPTER 6

How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?

Chapter abstract ...

...

PART III

One Class Random Forests

CHAPTER 7

One Class Splitting Criteria for Random Forests with Application to Anomaly Detection

Chapter abstract abstract ocrf

7.1 Introduction

...

PART IV

**Accuracy on Extreme
Regions**

CHAPTER 8

Learning the dependence structure of rare events: a non-asymptotic study

Chapter abstract This chapter presents the details relative to the introducing section 1.6. Assessing the probability of occurrence of extreme events is a crucial issue in various fields like finance, insurance, telecommunication or environmental sciences. In a multivariate framework, the tail dependence is characterized by the so-called *stable tail dependence function* (STDF). Learning this structure is the keystone of multivariate extremes. Although extensive studies have proved consistency and asymptotic normality for the empirical version of the STDF, non-asymptotic bounds are still missing. The main purpose of this paper is to fill this gap. Taking advantage of adapted VC-type concentration inequalities, upper bounds are derived with expected rate of convergence in $O(k^{-1/2})$. The concentration tools involved in this analysis rely on a more general study of maximal deviations in low probability regions, and thus directly apply to the classification of extreme data. The material of this chapter is based on previous work published in Goix et al. (2015b).

8.1 Introduction

As a first go, we briefly recall the preliminaries of Section 1.6.

To introduce the stable tail dependence function, suppose we want to manage the risk of a portfolio containing d different assets, $\mathbf{X} = (X_1, \dots, X_d)$. We want to evaluate the probability of events of the kind $\{X_1 \geq x_1 \text{ or } \dots \text{ or } X_d \geq x_d\}$, for large multivariate thresholds $\mathbf{x} = (x_1, \dots, x_d)$.

EVT shows that under not too strong condition on the regularity of the underlying tail distribution, for large enough thresholds, (see Section 8.2 for details)

$$\mathbb{P}\{X_1 \geq x_1 \text{ or } \dots \text{ or } X_d \geq x_d\} \simeq l(p_1, \dots, p_d),$$

where l is the *stable tail dependence function* and the p_j 's are the marginal exceedance probabilities, $p_j = \mathbb{P}(X_j \geq x_j)$. Thus, the functional l characterizes the *dependence* among extremes. The *joint* distribution (over large thresholds) can thus be recovered from the knowledge of the marginal distributions together with the STDF l . In practice, l can be learned from ‘moderately

extreme' data, typically the k 'largest' ones among a sample of size n , with $k \ll n$. Recovering the p_j 's can be easily done using univariate EVT modelling introduced in Section 4.1. However, in the multivariate case, there is no finite-dimensional parametrization of the dependence structure. The latter is characterized by the *stable tail dependence function* (STDF) l . Estimating this functional is thus one of the main issues in multivariate EVT. Asymptotic properties of the empirical STDF have been widely studied, see Huang (1992), Drees & Huang (1998), Embrechts et al. (2000) and de Haan & Ferreira (2006) for the bivariate case, and Qi (1997), Einmahl et al. (2012) for the general multivariate case under smoothness assumptions.

However, to the best of our knowledge, no bounds exist on the finite sample error. It is precisely the purpose of this paper to derive such non-asymptotic bounds. Our results do not require any assumption other than the existence of the STDF. The main idea is as follows. The empirical estimator is based on the empirical measure of 'extreme' regions, which are hit only with low probability. It is thus enough to bound maximal deviations on such low probability regions. The key consists in choosing an adaptive VC class, which only covers the latter regions, and on the other hand, to derive VC-type inequalities that incorporate p , the probability of hitting the class at all.

The structure of this chapter is as follows. The whys and wherefores of EVT and the STDF are explained in Section 8.2. In Section 8.3, concentration tools which rely on the general study of maximal deviations in low probability regions are introduced, with an immediate application to the framework of classification (Remark 8.5). The main result of this contribution, a non-asymptotic bound on the convergence of the empirical STDF, is derived in Section 8.4. Section 8.5 concludes.

8.2 Background on the stable tail dependence function

In the multivariate case, it is mathematically very convenient to decompose the joint distribution of $\mathbf{X} = (X^1, \dots, X^d)$ into the margins on the one hand, and the dependence structure on the other hand. In particular, handling uniform margins is very helpful when it comes to establishing upper bounds on the deviations between empirical and mean measures. Define thus standardized variables $U^j = 1 - F_j(X^j)$, where F_j is the marginal distribution function of X^j , and $\mathbf{U} = (U^1, \dots, U^d)$. Knowledge of the F_j 's and of the joint distribution of \mathbf{U} allows to recover that of \mathbf{X} , since $\mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d) = \mathbb{P}(U^1 \geq 1 - F_1(x_1), \dots, U^d \geq 1 - F_d(x_d))$. With these notations, under the fairly general assumption, namely, standard multivariate regular variation of standardized variables (4.5), equivalent to (4.11), there exists a limit measure Λ on $[0, \infty]^d \setminus \{\infty\}$ (called the *exponent measure*) such that

$$\lim_{t \rightarrow 0} t^{-1} \mathbb{P} \left[U^1 \leq t x_1 \text{ or } \dots \text{ or } U^d \leq t x_d \right] = \Lambda[\mathbf{x}, \infty]^c := l(\mathbf{x}) . \quad (x_j \in [0, \infty], \mathbf{x} \neq \infty) \quad (8.1)$$

Notice that no assumption is made about the marginal distributions, so that our framework allows non-standard regular variation, or even no regular variation at all of the original data \mathbf{X} (for more details see *e.g.* Resnick (2007), th. 6.5 or Resnick (1987), prop. 5.10.). The functional l in the limit in (8.1) is called the *stable tail dependence function*. In the remainder of this chapter, the only assumption is the existence of a limit in (8.1), *i.e.*, the existence of the STDF – or equivalently conditions (4.5) or (4.11) in the background section 4.2 on multivariate EVT.

We emphasize that the knowledge of both l and the margins gives access to the probability of hitting ‘extreme’ regions of the kind $[\mathbf{0}, \mathbf{x}]^c$, for ‘large’ thresholds $\mathbf{x} = (x_1, \dots, x_d)$ (*i.e.* such that for some $j \leq d$, $1 - F_j(x_j)$ is a $O(t)$ for some small t). Indeed, in such a case,

$$\begin{aligned} \mathbb{P}(X^1 > x_1 \text{ or } \dots \text{ or } X^d > x_d) &= \mathbb{P}\left(\bigcup_{j=1}^d (1 - F_j)(X^j) \leq (1 - F_j)(x_j)\right) \\ &= t \left\{ \frac{1}{t} \mathbb{P}\left(\bigcup_{j=1}^d U^j \leq t \left\lceil \frac{(1 - F_j)(x_j)}{t} \right\rceil\right) \right\} \\ &\underset{t \rightarrow 0}{\sim} t l\left(t^{-1}(1 - F_1)(x_1), \dots, t^{-1}(1 - F_d)(x_d)\right) \\ &= l\left((1 - F_1)(x_1), \dots, (1 - F_d)(x_d)\right) \end{aligned}$$

where the last equality follows from the homogeneity of l . This underlines the utmost importance of estimating the STDF and by extension stating non-asymptotic bounds on this convergence.

Any stable tail dependence function $l(\cdot)$ is in fact a norm, (see Falk et al. (1994), p179) and satisfies

$$\max\{x_1, \dots, x_n\} \leq l(\mathbf{x}) \leq x_1 + \dots + x_d,$$

where the lower bound is attained if \mathbf{X} is perfectly tail dependent (extremes of univariate marginals always occur simultaneously), and the upper bound in case of tail independence or asymptotic independence (extremes of univariate marginals never occur simultaneously). We refer to Falk et al. (1994) for more details and properties on the STDF.

8.3 A VC-type inequality adapted to the study of low probability regions

Classical VC inequalities aim at bounding the deviation of empirical from theoretical quantities on relatively simple classes of sets, called VC classes. These classes typically cover the support of the underlying distribution. However, when dealing with rare events, it is of great interest to have such bounds on a class of sets which only covers a small probability region and thus contains (very) few observations. This yields sharper bounds, since only differences between

very small quantities are involved. The starting point of this analysis is the following VC-inequality stated below.

Theorem 8.1. *Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. realizations of a r.v. \mathbf{X} , a VC-class \mathcal{A} with VC-dimension $V_{\mathcal{A}}$ and shattering coefficient (or growth function) $S_{\mathcal{A}}(n)$. Consider the class union $\mathbb{A} = \cup_{A \in \mathcal{A}} A$, and let $p = \mathbb{P}(\mathbf{X} \in \mathbb{A})$. Then there is an absolute constant C such that for all $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}[\mathbf{X} \in A] - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| \leq C \left[\sqrt{p} \sqrt{\frac{V_{\mathcal{A}}}{n} \log \frac{1}{\delta}} + \frac{1}{n} \log \frac{1}{\delta} \right]. \quad (8.2)$$

Proof. (sketch of) Details of the proof are deferred to the appendix section. We use a Bernstein-type concentration inequality (McDiarmid (1998)) that we apply to the general functional

$$f(\mathbf{X}_{1:n}) = \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right|,$$

where $\mathbf{X}_{1:n}$ denotes the sample $(\mathbf{X}_1, \dots, \mathbf{X}_n)$. The inequality in McDiarmid (1998) involves the variance of the r.v. $f(\mathbf{X}_1, \dots, \mathbf{X}_k, x_{k+1}, \dots, x_n) - f(\mathbf{X}_1, \dots, \mathbf{X}_{k-1}, x_k, \dots, x_n)$, which can easily be bounded in our setting. We obtain

$$\mathbb{P}[f(\mathbf{X}_{1:n}) - \mathbb{E}f(\mathbf{X}_{1:n}) \geq t] \leq e^{-\frac{nt^2}{2q + \frac{2t}{3}}}, \quad (8.3)$$

where the quantity $q = \mathbb{E}(\sup_{A \in \mathcal{A}} |\mathbb{1}_{\mathbf{X}' \in A} - \mathbb{1}_{\mathbf{X} \in A}|)$ (with \mathbf{X}' an independent copy of \mathbf{X}) is a measure of the complexity of the class \mathcal{A} with respect to the distribution of \mathbf{X} . It leads to high probability bounds on $f(\mathbf{X}_{1:n})$ of the form $\mathbb{E}f(\mathbf{X}_{1:n}) + \frac{1}{n} \log(1/\delta) + \sqrt{\frac{2q}{n} \log(1/\delta)}$ instead of the standard Hoeffding-type bound $\mathbb{E}f(\mathbf{X}_{1:n}) + \sqrt{\frac{1}{n} \log(1/\delta)}$. It is then easy to see that $q \leq 2 \sup_{A \in \mathcal{A}} \mathbb{P}(\mathbf{X} \in A) \leq 2p$. Finally, an upper bound on $\mathbb{E}f(\mathbf{X}_{1:n})$ is obtained by introducing re-normalized Rademacher averages

$$\mathcal{R}_{n,p} = \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right|.$$

which are then proved to be of order $O(\sqrt{\frac{V_{\mathcal{A}}}{pn}})$, so that $\mathbb{E}(f(\mathbf{X}_{1:n})) \leq C \sqrt{\frac{V_{\mathcal{A}}}{pn}}$. \square

Remark 8.2. (COMPARISON WITH EXISTING BOUNDS) The following re-normalized VC-inequality due to Vapnik and Chervonenkis (see Vapnik & Chervonenkis (1974), Anthony & Shawe-Taylor (1993) or Bousquet et al. (2004), Thm 7),

$$\sup_{A \in \mathcal{A}} \left| \frac{\mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A}}{\sqrt{\mathbb{P}(\mathbf{X} \in A)}} \right| \leq 2 \sqrt{\frac{\log S_{\mathcal{A}}(2n) + \log \frac{4}{\delta}}{n}}, \quad (8.4)$$

which holds under the same conditions as Theorem 8.1, allows to derive a bound similar to (8.2), but with an additional $\log n$ factor. Indeed, it is known as Sauer's Lemma (see Bousquet et al.

(2004)-lemma 1 for instance) that for $n \geq V_{\mathcal{A}}$, $S_{\mathcal{A}}(n) \leq (\frac{en}{V_{\mathcal{A}}})^{V_{\mathcal{A}}}$. It is then easy to see from (8.4) that:

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| \leq 2 \sqrt{\sup_{A \in \mathcal{A}} \mathbb{P}(\mathbf{X} \in A)} \sqrt{\frac{V_{\mathcal{A}} \log \frac{2en}{V_{\mathcal{A}}} + \log \frac{4}{\delta}}{n}}.$$

Introduce the union \mathbb{A} of all sets in the considered VC class, $\mathbb{A} = \cup_{A \in \mathcal{A}} A$, and let $p = \mathbb{P}(\mathbf{X} \in \mathbb{A})$. Then, the previous bound immediately yields

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| \leq 2\sqrt{p} \sqrt{\frac{V_{\mathcal{A}} \log \frac{2en}{V_{\mathcal{A}}} + \log \frac{4}{\delta}}{n}}.$$

Remark 8.3. (SIMPLER BOUND) If we assume furthermore that $\delta \geq e^{-np}$, then we have:

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| \leq C\sqrt{p} \sqrt{\frac{V_{\mathcal{A}}}{n} \log \frac{1}{\delta}}.$$

Remark 8.4. (INTERPRETATION) Inequality (8.2) can be seen as an interpolation between the best case (small p) where the rate of convergence is $O(1/n)$, and the worst case (large p) where the rate is $O(1/\sqrt{n})$. An alternative interpretation is as follows: divide both sides of (8.2) by p , so that the left hand side becomes a supremum of conditional probabilities upon belonging to the union class \mathbb{A} , $\{\mathbb{P}(\mathbf{X} \in A | \mathbf{X} \in \mathbb{A})\}_{A \in \mathcal{A}}$. Then the upper bound is proportional to $\epsilon(np, \delta)$ where $\epsilon(n, \delta) := \sqrt{\frac{V_{\mathcal{A}}}{n} \log \frac{1}{\delta}} + \frac{1}{n} \log \frac{1}{\delta}$ is a classical VC-bound; np is in fact the expected number of observations involved in (8.2), and can thus be viewed as the effective sample size.

Remark 8.5. (CLASSIFICATION OF EXTREMES) A key issue in the prediction framework is to find upper bounds for the maximal deviation $\sup_{g \in \mathcal{G}} |L_n(g) - L(g)|$, where $L(g) = \mathbb{P}(g(\mathbf{X}) \neq Y)$ is the risk of the classifier $g : \mathcal{X} \rightarrow \{-1, 1\}$, associated with the r.v. $(\mathbf{X}, Y) \in \mathbb{R}^d \times \{-1, 1\}$. $L_n(g) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{g(\mathbf{X}_i) \neq Y_i\}$ is the empirical risk based on a training dataset $\{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$. Strong upper bounds on $\sup_{g \in \mathcal{G}} |L_n(g) - L(g)|$ ensure the accuracy of the empirical risk minimizer $g_n := \arg \min_{g \in \mathcal{G}} L_n(g)$.

In a wide variety of applications (e.g. Finance, Insurance, Networks), it is of crucial importance to predict the system response Y when the input variable \mathbf{X} takes extreme values, corresponding to shocks on the underlying mechanism. In such a case, the risk of a prediction rule $g(\mathbf{X})$ should be defined by integrating the loss function $L(g)$ with respect to the conditional joint distribution of the pair (\mathbf{X}, Y) given \mathbf{X} is extreme. For instance, consider the event $\{\|\mathbf{X}\| \geq t_\alpha\}$ where t_α is the $(1 - \alpha)^{th}$ quantile of $\|\mathbf{X}\|$ for a small α . To investigate the accuracy of a classifier g given $\{\|\mathbf{X}\| \geq t_\alpha\}$, introduce

$$L_\alpha(g) := \frac{1}{\alpha} \mathbb{P}(Y \neq g(\mathbf{X}), \|\mathbf{X}\| > t_\alpha) = \mathbb{P}(Y \neq g(\mathbf{X}) \mid \|\mathbf{X}\| \geq t_\alpha),$$

and its empirical counterpart

$$L_{\alpha,n}(g) := \frac{1}{n\alpha} \sum_{i=1}^n \mathbb{I}_{\{Y_i \neq g(\mathbf{X}_i), \|\mathbf{X}_i\| > \|\mathbf{X}_{(\lfloor n\alpha \rfloor)}\|\}},$$

where $\|\mathbf{X}_{(1)}\| \geq \dots \geq \|\mathbf{X}_{(n)}\|$ are the order statistics of $\|\mathbf{X}\|$. Then as an application of Theorem 8.1 with $\mathcal{A} = \{(\mathbf{x}, y), g(\mathbf{x}) \neq y, \|\mathbf{x}\| > t_\alpha\}$, $g \in \mathcal{G}$, we have :

$$\sup_{g \in \mathcal{G}} \left| \widehat{L}_{\alpha,n}(g) - L_\alpha(g) \right| \leq C \left[\sqrt{\frac{V_{\mathcal{G}}}{n\alpha} \log \frac{1}{\delta}} + \frac{1}{n\alpha} \log \frac{1}{\delta} \right]. \quad (8.5)$$

We refer to the appendix for more details. Again the obtained rate by empirical risk minimization meets our expectations (see remark 8.4), insofar as α is the fraction of the dataset involved in the empirical risk $L_{\alpha,n}$. We point out that α may typically depend on n , $\alpha = \alpha_n \rightarrow 0$. In this context a direct use of the standard version of the VC inequality would lead to a rate of order $1/(\alpha_n \sqrt{n})$, which may not vanish as $n \rightarrow +\infty$ and even go to infinity if α_n decays to 0 faster than $1/\sqrt{n}$.

Let us point out that rare events may be chosen more general than $\{\|\mathbf{X}\| > t_\alpha\}$, say $\{\mathbf{X} \in Q\}$ with unknown probability $q = \mathbb{P}(\{\mathbf{X} \in Q\})$. The previous result still applies with $\widetilde{L}_Q(g) := \mathbb{P}(Y \neq g(\mathbf{X}), \mathbf{X} \in Q)$ and $\widetilde{L}_{Q,n}(g) := \mathbb{P}_n(Y \neq g(\mathbf{X}), \mathbf{X} \in Q)$; then the obtained upper bound on $\sup_{g \in \mathcal{G}} \frac{1}{q} \left| \widetilde{L}_Q(g) - \widetilde{L}_{Q,n}(g) \right|$ is of order $O(1/\sqrt{qn})$.

Similar results can be established for the problem of *distribution-free regression*, when the error of any predictive rule $f(\mathbf{x})$ is measured by the conditional mean squared error $\mathbb{E}[(Z - f(\mathbf{X}))^2 \mid Z > q_{\alpha_n}]$, denoting by Z the real-valued output variable to be predicted from \mathbf{X} and by q_α its quantile at level $1 - \alpha$.

8.4 A bound on the STDF

Let us place ourselves in the multivariate extreme framework introduced in Section 8.1: Consider a random variable $\mathbf{X} = (X^1, \dots, X^d)$ in \mathbb{R}^d with distribution function F and marginal distribution functions F_1, \dots, F_d . Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be an *i.i.d.* sample distributed as \mathbf{X} . In the subsequent analysis, the only assumption is the existence of the STDF defined in (8.1) and the margins F_j are supposed to be unknown. The definition of l may be recast as

$$l(\mathbf{x}) := \lim_{t \rightarrow 0} t^{-1} \widetilde{F}(t\mathbf{x}) \quad (8.6)$$

with $\widetilde{F}(\mathbf{x}) = (1 - F)((1 - F_1)^{\leftarrow}(x_1), \dots, (1 - F_d)^{\leftarrow}(x_d))$. Here the notation $(1 - F_j)^{\leftarrow}(x_j)$ denotes the quantity $\sup\{y : 1 - F_j(y) \geq x_j\}$. Notice that, in terms of standardized variables U^j , $\widetilde{F}(\mathbf{x}) = \mathbb{P}\left(\bigcup_{j=1}^d \{U^j \leq x_j\}\right) = \mathbb{P}(\mathbf{U} \in [\mathbf{x}, \infty]^c)$.

Let $k = k(n)$ be a sequence of positive integers such that $k \rightarrow \infty$ and $k = o(n)$ as $n \rightarrow \infty$. A natural estimator of l is its empirical version defined as follows, see Huang (1992), Qi (1997),

Drees & Huang (1998), Einmahl et al. (2006):

$$l_n(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{X_i^1 \geq X_{(n-\lfloor kx_1 \rfloor + 1)}^1 \text{ or } \dots \text{ or } X_i^d \geq X_{(n-\lfloor kx_d \rfloor + 1)}^d\}}, \quad (8.7)$$

The expression is indeed suggested by the definition of l in (8.6), with all distribution functions and univariate quantiles replaced by their empirical counterparts, and with t replaced by k/n . Extensive studies have proved consistency and asymptotic normality of this nonparametric estimator of l , see Huang (1992), Drees & Huang (1998) and de Haan & 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 .

To our best knowledge, there is no established non-asymptotic bound on the maximal deviation $\sup_{0 \leq \mathbf{x} \leq T} |l_n(\mathbf{x}) - l(\mathbf{x})|$. It is the purpose of the remainder of this section to derive such a bound, without any smoothness condition on l .

First, Theorem 8.1 needs adaptation to a particular setting: introduce a random vector $\mathbf{Z} = (Z^1, \dots, Z^d)$ with uniform margins, *i.e.*, for every $j = 1, \dots, d$, the variable Z^j is uniform on $[0, 1]$. Consider the class

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

This is a VC-class of VC-dimension d , as proved in Devroye et al. (1996), Theorem 13.8, for its complementary class $\{[\mathbf{x}, \infty[, \mathbf{x} > 0\}$. In this context, the union class \mathbb{A} has mass $p \leq dT \frac{k}{n}$ since

$$\mathbb{P}(\mathbf{Z} \in \mathbb{A}) = \mathbb{P} \left[\mathbf{Z} \in \left(\left[\frac{k}{n} T, \infty \right]^d \right)^c \right] = \mathbb{P} \left[\bigcup_{j=1..d} \mathbf{Z}^j < \frac{k}{n} T \right] \leq \sum_{j=1}^d \mathbb{P} \left[\mathbf{Z}^j < \frac{k}{n} T \right]$$

Consider the measures $C_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{Z_i \in \cdot\}}$ and $C(\mathbf{x}) = \mathbb{P}(\mathbf{Z} \in \cdot)$. As a direct consequence of Theorem 8.1 the following inequality holds true with probability at least $1 - \delta$,

$$\sup_{0 \leq \mathbf{x} \leq T} \frac{n}{k} \left| C_n \left(\frac{k}{n} [\mathbf{x}, \infty]^c \right) - C \left(\frac{k}{n} [\mathbf{x}, \infty]^c \right) \right| \leq Cd \left(\sqrt{\frac{T}{k} \log \frac{1}{\delta}} + \frac{1}{k} \log \frac{1}{\delta} \right).$$

If we assume furthermore that $\delta \geq e^{-k}$, then we have

$$\sup_{0 \leq \mathbf{x} \leq T} \frac{n}{k} \left| C_n \left(\frac{k}{n} [\mathbf{x}, \infty]^c \right) - C \left(\frac{k}{n} [\mathbf{x}, \infty]^c \right) \right| \leq Cd \sqrt{\frac{T}{k} \log \frac{1}{\delta}}. \quad (8.8)$$

Inequality (8.8) is the cornerstone of the following theorem, which is the main result of this contribution. In the sequel, we consider a sequence $k(n)$ of integers such that $k = o(n)$ and

$k(n) \rightarrow \infty$. For notational convenience, we often drop the dependence in n and simply write k instead of $k(n)$.

Theorem 8.6. *Let T be a positive number such that $T \geq \frac{7}{2}(\frac{\log d}{k} + 1)$, and δ such that $\delta \geq e^{-k}$. Then there is an absolute constant C such that for each $n > 0$, with probability at least $1 - \delta$:*

$$\sup_{0 \leq \mathbf{x} \leq T} |l_n(\mathbf{x}) - l(\mathbf{x})| \leq Cd \sqrt{\frac{T}{k} \log \frac{d+3}{\delta}} + \sup_{0 \leq \mathbf{x} \leq 2T} \left| \frac{n}{k} \tilde{F}\left(\frac{k}{n} \mathbf{x}\right) - l(\mathbf{x}) \right| \quad (8.9)$$

The second term on the right hand side of (8.9) is a bias term which depends on the discrepancy between the left hand side and the limit in (8.1) or (8.6) 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. Considering classical inequalities in empirical process theory such as VC-bounds, it is thus no surprise to obtain one in $O(1/\sqrt{k})$. Too large values of k tend to yield a large bias, whereas too small values of k yield a large variance. For a more detailed discussion on the choice of k we recommend Einmahl et al. (2009).

The proof of Theorem 8.6 follows the same lines as in Qi (1997). For unidimensional random variables Y_1, \dots, Y_n , let us denote by $Y_{(1)} \leq \dots \leq Y_{(n)}$ their order statistics. Define then the empirical version \tilde{F}_n of \tilde{F} (introduced in (8.6)) as

$$\tilde{F}_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i^1 \leq x_1 \text{ or } \dots \text{ or } U_i^d \leq x_d\}},$$

so that $\frac{n}{k} \tilde{F}_n(\frac{k}{n} \mathbf{x}) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{U_i^1 \leq \frac{k}{n} x_1 \text{ or } \dots \text{ or } U_i^d \leq \frac{k}{n} x_d\}}$. Notice that the U_i^j 's are not observable (since F_j is unknown). In fact, \tilde{F}_n will be used as a substitute for l_n allowing to handle uniform variables. The following lemmas make this point explicit.

Lemma 8.7 (Link between l_n and \tilde{F}_n). *The empirical version of \tilde{F} and that of l are related via*

$$l_n(\mathbf{x}) = \frac{n}{k} \tilde{F}_n(U_{(\lfloor kx_1 \rfloor)}^1, \dots, U_{(\lfloor kx_d \rfloor)}^d).$$

Proof. Consider the definition of l_n in (8.7), and note that for $j = 1, \dots, d$,

$$\begin{aligned} X_i^j \geq X_{(n - \lfloor kx_j \rfloor + 1)}^j &\Leftrightarrow \text{rank}(X_i^j) \geq n - \lfloor kx_j \rfloor + 1 \\ &\Leftrightarrow \text{rank}(F_j(X_i^j)) \geq n - \lfloor kx_j \rfloor + 1 \\ &\Leftrightarrow \text{rank}(1 - F_j(X_i^j)) \leq \lfloor kx_j \rfloor \\ &\Leftrightarrow U_i^j \leq U_{(\lfloor kx_j \rfloor)}^j, \end{aligned}$$

so that $l_n(\mathbf{x}) = \frac{1}{k} \sum_{j=1}^n \mathbb{1}_{\{U_j^1 \leq U_{(\lfloor kx_1 \rfloor)}^1 \text{ or } \dots \text{ or } U_j^d \leq U_{(\lfloor kx_d \rfloor)}^d\}}$. □

Lemma 8.8 (Uniform bound on \tilde{F}_n 's deviations). *For any finite $T > 0$, and $\delta \geq e^{-k}$, with probability at least $1 - \delta$, the deviation of \tilde{F}_n from \tilde{F} is uniformly bounded:*

$$\sup_{0 \leq \mathbf{x} \leq T} \left| \frac{n}{k} \tilde{F}_n\left(\frac{k}{n} \mathbf{x}\right) - \frac{n}{k} \tilde{F}\left(\frac{k}{n} \mathbf{x}\right) \right| \leq Cd \sqrt{\frac{T}{k} \log \frac{1}{\delta}}$$

Proof. Notice that

$$\sup_{0 \leq \mathbf{x} \leq T} \left| \frac{n}{k} \tilde{F}_n\left(\frac{k}{n} \mathbf{x}\right) - \frac{n}{k} \tilde{F}\left(\frac{k}{n} \mathbf{x}\right) \right| = \frac{n}{k} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\mathbf{U}_i \in \frac{k}{n} [\mathbf{x}, \infty]^c\}} - \mathbb{P} \left[\mathbf{U} \in \frac{k}{n} [\mathbf{x}, \infty]^c \right] \right|,$$

and apply inequality (8.8). □

Lemma 8.9 (Bound on the order statistics of \mathbf{U}). *Let $\delta \geq e^{-k}$. For any finite positive number $T > 0$ such that $T \geq 7/2((\log d)/k + 1)$, we have with probability greater than $1 - \delta$,*

$$\forall 1 \leq j \leq d, \quad \frac{n}{k} U_{(\lfloor kT \rfloor)}^j \leq 2T, \quad (8.10)$$

and with probability greater than $1 - (d + 1)\delta$,

$$\max_{1 \leq j \leq d} \sup_{0 \leq x_j \leq T} \left| \frac{\lfloor kx_j \rfloor}{k} - \frac{n}{k} U_{(\lfloor kx_j \rfloor)}^j \right| \leq C \sqrt{\frac{T}{k} \log \frac{1}{\delta}}.$$

Proof. Notice that $\sup_{[0, T]} \frac{n}{k} U_{(\lfloor k \cdot \rfloor)}^j = \frac{n}{k} U_{(\lfloor kT \rfloor)}^j$ and let $\Gamma_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i^j \leq t\}}$. It then straightforward to see that

$$\frac{n}{k} U_{(\lfloor kT \rfloor)}^j \leq 2T \Leftrightarrow \Gamma_n\left(\frac{k}{n} 2T\right) \geq \frac{\lfloor kT \rfloor}{n}$$

so that

$$\mathbb{P} \left(\frac{n}{k} U_{(\lfloor kT \rfloor)}^j > 2T \right) \leq \mathbb{P} \left(\sup_{\frac{2kT}{n} \leq t \leq 1} \frac{t}{\Gamma_n(t)} > 2 \right).$$

Using Wellner (1978), Lemma 1-(ii) (we use the fact that, with the notations of this reference, $h(1/2) \geq 1/7$), we obtain

$$\mathbb{P} \left(\frac{n}{k} U_{(\lfloor kT \rfloor)}^j > 2T \right) \leq e^{-\frac{2kT}{7}},$$

and thus

$$\mathbb{P} \left(\exists j, \frac{n}{k} U_{(\lfloor kT \rfloor)}^j > 2T \right) \leq de^{-\frac{2kT}{7}} \leq e^{-k} \leq \delta$$

as required in (8.10). Yet,

$$\begin{aligned}
\sup_{0 \leq x_j \leq T} \left| \frac{\lfloor kx_j \rfloor}{k} - \frac{n}{k} U_{(\lfloor kx_j \rfloor)}^j \right| &= \sup_{0 \leq x_j \leq T} \left| \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{U_i^j \leq U_{(\lfloor kx_j \rfloor)}^j\}} - \frac{n}{k} U_{(\lfloor kx_j \rfloor)}^j \right| \\
&= \frac{n}{k} \sup_{0 \leq x_j \leq T} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i^j \leq U_{(\lfloor kx_j \rfloor)}^j\}} - \mathbb{P} \left[U_1^j \leq U_{(\lfloor kx_j \rfloor)}^j \right] \right| \\
&= \sup_{0 \leq x_j \leq T} \Theta_j \left(\frac{n}{k} U_{(\lfloor kx_j \rfloor)}^j \right),
\end{aligned}$$

where $\Theta_j(y) = \frac{n}{k} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i^j \leq \frac{k}{n}y\}} - \mathbb{P} \left[U_1^j \leq \frac{k}{n}y \right] \right|$. Then, by (8.10), with probability greater than $1 - \delta$,

$$\max_{1 \leq j \leq d} \sup_{0 \leq x_j \leq T} \left| \frac{\lfloor kx_j \rfloor}{k} - \frac{n}{k} U_{(\lfloor kx_j \rfloor)}^j \right| \leq \max_{1 \leq j \leq d} \sup_{0 \leq y \leq 2T} \Theta_j(y)$$

and from (8.8), each term $\sup_{0 \leq y \leq 2T} \Theta_j(y)$ is bounded by $C \sqrt{\frac{T}{k} \log \frac{1}{\delta}}$ (with probability $1 - \delta$). In the end, with probability greater than $1 - (d + 1)\delta$:

$$\max_{1 \leq j \leq d} \sup_{0 \leq y \leq 2T} \Theta_j(y) \leq C \sqrt{\frac{T}{k} \log \frac{1}{\delta}},$$

which is the desired inequality □

We may now proceed with the proof of Theorem 8.6. First of all, noticing that $\tilde{F}(t\mathbf{x})$ is non-decreasing in x_j for every l and that $l(\mathbf{x})$ is non-decreasing and continuous (thus uniformly continuous on $[0, T]^d$), from (8.6) it is easy to prove by subdividing $[0, T]^d$ (see Qi (1997) p.174 for details) that

$$\sup_{0 \leq \mathbf{x} \leq T} \left| \frac{1}{t} \tilde{F}(t\mathbf{x}) - l(\mathbf{x}) \right| \rightarrow 0 \quad \text{as } t \rightarrow 0. \quad (8.11)$$

Using Lemma 8.7, we can write :

$$\begin{aligned}
\sup_{0 \leq \mathbf{x} \leq T} |l_n(\mathbf{x}) - l(\mathbf{x})| &= \sup_{0 \leq \mathbf{x} \leq T} \left| \frac{n}{k} \tilde{F}_n \left(U_{(\lfloor kx_1 \rfloor)}^1, \dots, U_{(\lfloor kx_d \rfloor)}^d \right) - l(\mathbf{x}) \right| \\
&\leq \sup_{0 \leq \mathbf{x} \leq T} \left| \frac{n}{k} \tilde{F}_n \left(U_{(\lfloor kx_1 \rfloor)}^1, \dots, U_{(\lfloor kx_d \rfloor)}^d \right) - \frac{n}{k} \tilde{F} \left(U_{(\lfloor kx_1 \rfloor)}^1, \dots, U_{(\lfloor kx_d \rfloor)}^d \right) \right| \\
&\quad + \sup_{0 \leq \mathbf{x} \leq T} \left| \frac{n}{k} \tilde{F} \left(U_{(\lfloor kx_1 \rfloor)}^1, \dots, U_{(\lfloor kx_d \rfloor)}^d \right) - l \left(\frac{n}{k} U_{(\lfloor kx_1 \rfloor)}^1, \dots, \frac{n}{k} U_{(\lfloor kx_d \rfloor)}^d \right) \right| \\
&\quad + \sup_{0 \leq \mathbf{x} \leq T} \left| l \left(\frac{n}{k} U_{(\lfloor kx_1 \rfloor)}^1, \dots, \frac{n}{k} U_{(\lfloor kx_d \rfloor)}^d \right) - l(\mathbf{x}) \right| \\
&=: \Lambda(n) + \Xi(n) + \Upsilon(n).
\end{aligned}$$

Now, by (8.10) we have with probability greater than $1 - \delta$:

$$\Lambda(n) \leq \sup_{0 \leq \mathbf{x} \leq 2T} \left| \frac{n}{k} \tilde{F}_n\left(\frac{k}{n} \mathbf{x}\right) - \frac{n}{k} \tilde{F}\left(\frac{k}{n} \mathbf{x}\right) \right|$$

and by Lemma 8.8,

$$\Lambda(n) \leq Cd \sqrt{\frac{2T}{k} \log \frac{1}{\delta}}$$

with probability at least $1 - 2\delta$. Similarly,

$$\Xi(n) \leq \sup_{0 \leq \mathbf{x} \leq 2T} \left| \frac{n}{k} \tilde{F}\left(\frac{k}{n} \mathbf{x}\right) - \frac{n}{k} l\left(\frac{k}{n} \mathbf{x}\right) \right| = \sup_{0 \leq \mathbf{x} \leq 2T} \left| \frac{n}{k} \tilde{F}\left(\frac{k}{n} \mathbf{x}\right) - l(\mathbf{x}) \right| \rightarrow 0 \quad (\text{bias term})$$

by virtue of (8.11). Concerning $\Upsilon(n)$, we have :

$$\begin{aligned} \Upsilon(n) &\leq \sup_{0 \leq \mathbf{x} \leq T} \left| l\left(\frac{n}{k} U_{(\lfloor kx_1 \rfloor)}^1, \dots, \frac{n}{k} U_{(\lfloor kx_d \rfloor)}^d\right) - l\left(\frac{\lfloor kx_1 \rfloor}{k}, \dots, \frac{\lfloor kx_d \rfloor}{k}\right) \right| \\ &\quad + \sup_{0 \leq \mathbf{x} \leq T} \left| l\left(\frac{\lfloor kx_1 \rfloor}{k}, \dots, \frac{\lfloor kx_d \rfloor}{k}\right) - l(\mathbf{x}) \right| \\ &= \Upsilon_1(n) + \Upsilon_2(n) \end{aligned}$$

Recall that l is 1-Lipschitz on $[0, T]^d$ regarding to the $\|\cdot\|_1$ -norm, so that

$$\Upsilon_1(n) \leq \sup_{0 \leq \mathbf{x} \leq T} \sum_{l=1}^d \left| \frac{\lfloor kx_j \rfloor}{k} - \frac{n}{k} U_{(\lfloor kx_j \rfloor)}^j \right|$$

so that by Lemma 8.9, with probability greater than $1 - (d+1)\delta$:

$$\Upsilon_1(n) \leq Cd \sqrt{\frac{2T}{k} \log \frac{1}{\delta}}.$$

On the other hand, $\Upsilon_2(n) \leq \sup_{0 \leq \mathbf{x} \leq T} \sum_{l=1}^d \left| \frac{\lfloor kx_j \rfloor}{k} - x_j \right| \leq \frac{d}{k}$. Finally we get, for every $n > 0$, with probability at least $1 - (d+3)\delta$:

$$\begin{aligned} \sup_{0 \leq \mathbf{x} \leq T} |l_n(\mathbf{x}) - l(\mathbf{x})| &\leq \Lambda(n) + \Upsilon_1(n) + \Upsilon_2(n) + \Xi(n) \\ &\leq Cd \sqrt{\frac{2T}{k} \log \frac{1}{\delta}} + Cd \sqrt{\frac{2T}{k} \log \frac{1}{\delta}} + \frac{d}{k} + \sup_{0 \leq \mathbf{x} \leq 2T} \left| \tilde{F}(\mathbf{x}) - \frac{n}{k} l\left(\frac{k}{n} \mathbf{x}\right) \right| \\ &\leq C'd \sqrt{\frac{2T}{k} \log \frac{1}{\delta}} + \sup_{0 \leq \mathbf{x} \leq 2T} \left| \frac{n}{k} \tilde{F}\left(\frac{k}{n} \mathbf{x}\right) - l(\mathbf{x}) \right| \end{aligned}$$

8.5 Discussion

We provide a non-asymptotic bound of VC type controlling the error of the empirical version of the STDF. Our bound achieves the expected rate in $O(k^{-1/2}) + \text{bias}(k)$, where k is the number of (extreme) observations retained in the learning process. In practice the smaller k/n , the smaller the bias. Since no assumption is made on the underlying distribution, other than the existence of the STDF, it is not possible in our framework to control the bias explicitly. One option would be to make an additional hypothesis of ‘second order regular variation’ (see *e.g.* de Haan & Resnick, 1996). We made the choice of making as few assumptions as possible, however, since the bias term is separated from the ‘variance’ term, it is probably feasible to refine our result with more assumptions.

For the purpose of controlling the empirical STDF, we have adopted the more general framework of maximal deviations in low probability regions. The VC-type bounds adapted to low probability regions derived in Section 8.3 may directly be applied to a particular prediction context, namely where the objective is to learn a classifier (or a regressor) that has good properties on low probability regions. This may open the road to the study of classification of extremal observations, with immediate applications to the field of anomaly detection.

8.6 Proof of Theorem 8.1

Theorem 8.1 is actually a short version of Theorem 8.10 below:

Theorem 8.10 (Maximal deviations). *Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. realizations of a r.v. \mathbf{X} valued in \mathbb{R}^d , a VC-class \mathcal{A} , and denote by $\mathcal{R}_{n,p}$ the associated relative Rademacher average defined by*

$$\mathcal{R}_{n,p} = \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{np} \left| \sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right|. \quad (8.12)$$

Define the union $\mathbb{A} = \cup_{A \in \mathcal{A}} A$, and $p = \mathbb{P}(\mathbf{X} \in \mathbb{A})$. Fix $0 < \delta < 1$, then with probability at least $1 - \delta$,

$$\frac{1}{p} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| \leq 2\mathcal{R}_{n,p} + \frac{2}{3np} \log \frac{1}{\delta} + 2\sqrt{\frac{1}{np} \log \frac{1}{\delta}},$$

and there is a constant C independent of n, p, δ such that with probability greater than $1 - \delta$,

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| \leq C \left(\sqrt{p} \sqrt{\frac{V_A}{n} \log \frac{1}{\delta}} + \frac{1}{n} \log \frac{1}{\delta} \right).$$

If we assume furthermore that $\delta \geq e^{-np}$, then we both have:

$$\begin{aligned} \frac{1}{p} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| &\leq 2\mathcal{R}_{n,p} + 3\sqrt{\frac{1}{np} \log \frac{1}{\delta}} \\ \frac{1}{p} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| &\leq C\sqrt{\frac{V_{\mathcal{A}}}{np} \log \frac{1}{\delta}}. \end{aligned}$$

In the following, $\mathbf{X}_{1:n}$ denotes an *i.i.d.* sample $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ distributed as \mathbf{X} , a \mathbb{R}^d -valued random vector. The classical steps to prove VC inequalities consist in applying a concentration inequality to the function

$$f(\mathbf{X}_{1:n}) := \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right|, \quad (8.13)$$

and then establishing bounds on the expectation $\mathbb{E}f(\mathbf{X}_{1:n})$, using for instance Rademacher average. Here we follow the same lines, but applying a Bernstein type concentration inequality instead of the usual Hoeffding one, since the variance term in the bound involves the probability p to be in the union of the VC-class \mathcal{A} considered. We then introduce relative Rademacher averages instead of the conventional ones, to take into account p for bounding $\mathbb{E}f(\mathbf{X}_{1:n})$.

We need first to control the variability of the random variable $f(\mathbf{X}_{1:n})$ when fixing all but one marginal \mathbf{X}_i . For that purpose introduce the functional

$$h(\mathbf{x}_1, \dots, \mathbf{x}_k) = \mathbb{E}[f(\mathbf{X}_{1:n}) | \mathbf{X}_1 = \mathbf{x}_1, \dots, \mathbf{X}_k = \mathbf{x}_k] - \mathbb{E}[f(\mathbf{X}_{1:n}) | \mathbf{X}_1 = \mathbf{x}_1, \dots, \mathbf{X}_{k-1} = \mathbf{x}_{k-1}]$$

The *positive deviation* of $h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{X}_k)$ is defined by

$$dev^+(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \{h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{x})\},$$

and \maxdev^+ , the maximum of all positive deviations, by

$$\maxdev^+ = \sup_{\mathbf{x}_1, \dots, \mathbf{x}_{k-1}} \max_k dev^+(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}).$$

Finally, define \hat{v} , the *maximum sum of variances*, by

$$\hat{v} = \sup_{\mathbf{x}_1, \dots, \mathbf{x}_n} \sum_{k=1}^n \text{Var } h(\mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{X}_k).$$

We have now the tools to state an extension of the classical Bernstein inequality, which is proved in McDiarmid (1998).

Proposition 8.11. *Let $\mathbf{X}_{1:n} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ as above, and f any function $(\mathbb{R}^d)^n \rightarrow \mathbb{R}$. Let \maxdev^+ and \hat{v} the maximum sum of variances, both of which we assume to be finite, and let μ*

be the mean of $f(\mathbf{X}_{1:n})$. Then for any $t \geq 0$,

$$\mathbb{P}[f(\mathbf{X}_{1:n}) - \mu \geq t] \leq \exp\left(-\frac{t^2}{2\hat{v}(1 + \frac{\maxdev^+ t}{3\hat{v}})}\right).$$

Note that the term $\frac{\maxdev^+ t}{3\hat{v}}$ is view as an ‘error term’ and is often negligible. Let us apply this theorem to the specific function f defined in (8.13). Then the following lemma (which have been stated and proved in the Chapter 3, see Lemma 3.19) holds true:

Lemma 8.12. *In the situation of Proposition 8.11 with f as in (8.13), we have*

$$\maxdev^+ \leq \frac{1}{n} \text{ and } \hat{v} \leq \frac{q}{n},$$

where

$$q = \mathbb{E}\left(\sup_{A \in \mathcal{A}} |\mathbb{1}_{\mathbf{X}' \in A} - \mathbb{1}_{\mathbf{X} \in A}|\right) \leq 2\mathbb{E}\left(\sup_{A \in \mathcal{A}} |\mathbb{1}_{\mathbf{X}' \in A} \mathbb{1}_{\mathbf{X} \notin A}|\right), \quad (8.14)$$

with \mathbf{X}' an independent copy of \mathbf{X} .

As a consequence with Proposition 8.11 the following general inequality holds true:

$$\mathbb{P}[f(\mathbf{X}_{1:n}) - \mathbb{E}f(\mathbf{X}_{1:n}) \geq t] \leq e^{-\frac{nt^2}{2q + \frac{2t}{3}}} \quad (8.15)$$

where the quantity $q = \mathbb{E}(\sup_{A \in \mathcal{A}} |\mathbb{1}_{\mathbf{X}' \in A} - \mathbb{1}_{\mathbf{X} \in A}|)$ seems to be a central characteristic of the VC-class \mathcal{A} given the distribution \mathbf{X} . It may be interpreted as a measure of the complexity of the class \mathcal{A} with respect to the distribution of \mathbf{X} : how often the class \mathcal{A} is able to separate two independent realizations of \mathbf{X} .

Recall that the union class \mathbb{A} and its associated probability p are defined as $\mathbb{A} = \cup_{A \in \mathcal{A}} A$, and $p = \mathbb{P}(\mathbf{X} \in \mathbb{A})$. Noting that for all $A \in \mathcal{A}$, $\mathbb{1}_{\{\cdot \in A\}} \leq \mathbb{1}_{\{\cdot \in \mathbb{A}\}}$, it is then straightforward from (8.14) that $q \leq 2p$. As a consequence (8.15) holds true when changing q by $2p$. Let us now explicit the link between the expectation of f and the Rademacher average

$$\mathcal{R}_n = \mathbb{E} \sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \right|,$$

where $(\sigma_i)_{i \geq 1}$ is a Rademacher chaos independent of the \mathbf{X}_i ’s.

Lemma 8.13. *With this notations the following inequality holds true:*

$$\mathbb{E}f(\mathbf{X}_{1:n}) \leq 2\mathcal{R}_n$$

This lemma has been proved in Chapter 3 in Theorem 3.16. Combining (8.15) with Lemma 8.13 and the fact that $q \leq 2p$ gives:

$$\mathbb{P}[f(\mathbf{X}_{1:n}) - 2\mathcal{R}_n \geq t] \leq e^{-\frac{nt^2}{4p+\frac{2t}{3}}}. \quad (8.16)$$

Recall that the relative Rademacher average are defined in (8.12) as $\mathcal{R}_{n,p} = \mathcal{R}_n/p$. It is well-known that \mathcal{R}_n is of order $\mathcal{O}((V_{\mathcal{A}}/n)^{1/2})$, see Koltchinskii (2006) for instance. However, we hope a stronger bound than just $\mathcal{R}_{n,p} = \mathcal{O}(p^{-1}(V_{\mathcal{A}}/n)^{1/2})$ since $\frac{1}{np} |\sum_{i=1}^n \sigma_i \mathbb{1}_{\mathbf{X}_i \in A}|$ with $\mathbb{P}(\mathbf{X}_i \in \mathbb{A}) = p$ is expected to be like $\frac{1}{np} |\sum_{i=1}^{np} \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A}|$ with \mathbf{Y}_i such that $\mathbb{P}(\mathbf{Y}_i \in \mathbb{A}) = 1$. The result below confirms this heuristic:

Lemma 8.14. *The relative Rademacher average $\mathcal{R}_{n,p}$ is of order $\mathcal{O}(\sqrt{\frac{V_{\mathcal{A}}}{pn}})$.*

This lemma has been proved in Chapter 3, see Lemma 3.22.

Finally we obtain from (8.16) and Lemma 8.14 the following bound:

$$\mathbb{P}\left[\frac{1}{p} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| - 2\mathcal{R}_{n,p} > t\right] \leq e^{-\frac{np t^2}{4+\frac{2t}{3}}} \quad (8.17)$$

Solving $\exp\left[-\frac{np t^2}{4+\frac{2t}{3}}\right] = \delta$ with $t > 0$ leads to

$$t = \frac{1}{3np} \log \frac{1}{\delta} + \sqrt{\left(\frac{1}{3np} \log \frac{1}{\delta}\right)^2 + \frac{4}{np} \log \frac{1}{\delta}} := h(\delta)$$

so that

$$\mathbb{P}\left[\frac{1}{p} \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| - 2\mathcal{R}_{n,p} > h(\delta)\right] \leq \delta$$

Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ if $a, b \geq 0$, we have $h(\delta) < \frac{2}{3np} \log \frac{1}{\delta} + 2\sqrt{\frac{1}{np} \log \frac{1}{\delta}}$. In the case of $\delta \geq e^{-np}$, $\frac{2}{3np} \log \frac{1}{\delta} \leq \frac{2}{3}\sqrt{\frac{1}{np} \log \frac{1}{\delta}}$ so that $h(\delta) < 3\sqrt{\frac{1}{np} \log \frac{1}{\delta}}$. This ends the proof.

8.7 Note on Remark 8.5

To obtain the bound in (8.5), the following easy to show inequality is needed before applying Theorem 8.1 :

$$\begin{aligned} \sup_{g \in \mathcal{G}} |L_{\alpha,n}(g) - L_{\alpha}(g)| &\leq \frac{1}{\alpha} \left[\sup_{g \in \mathcal{G}} \left| \mathbb{P}(Y \neq g(\mathbf{X}), \|\mathbf{X}\| > t_{\alpha}) - \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{Y_i \neq g(\mathbf{X}_i), \|\mathbf{X}_i\| > t_{\alpha}\}} \right| \right. \\ &\quad \left. + \left| \mathbb{P}(\|\mathbf{X}\| > t_{\alpha}) - \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{\|\mathbf{X}_i\| > t_{\alpha}\}} \right| + \frac{1}{n} \right]. \blacksquare \end{aligned}$$

Note that the final objective would be to bound the quantity $\sup_{g \in \mathcal{G}} |L_\alpha(g) - L_\alpha(g_\alpha^*)|$, where g_α^* is a Bayes classifier for the problem at stake, *i.e.* a solution of the conditional risk minimization problem $\inf_{\{g \text{ measurable}\}} L_\alpha(g)$. Such a bound involves a bias term $\inf_{g \in \mathcal{G}} L_\alpha(g) - L_\alpha(g_\alpha^*)$, as in the classical setting. Further, it can be shown that the standard Bayes classifier $g^*(\mathbf{x}) := 2\mathbb{I}\{\eta(\mathbf{x}) > 1/2\} - 1$ (where $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$) is also a solution of the conditional risk minimization problem. Finally, the conditional bias $\inf_{g \in \mathcal{G}} L_\alpha(g) - L_\alpha(g_\alpha^*)$ can be expressed as $\frac{1}{\alpha} \inf_{g \in \mathcal{G}} \mathbb{E} [|2\eta(\mathbf{X}) - 1| \mathbb{1}_{g(\mathbf{X}) \neq g^*(\mathbf{X})} \mathbb{1}_{\|\mathbf{X}\| \geq t_\alpha}]$, to be compared with the standard bias $\inf_{g \in \mathcal{G}} \mathbb{E} [|2\eta(\mathbf{X}) - 1| \mathbb{1}_{g(\mathbf{X}) \neq g^*(\mathbf{X})}]$.

Chapter abstract This chapter presents the details relative to the introducing Section 1.7. Capturing the dependence structure of multivariate extreme events is a major concern in many fields involving the management of risks stemming from multiple sources, *e.g.* portfolio monitoring, insurance, environmental risk management and anomaly detection. One convenient (nonparametric) characterization of extreme dependence in the framework of multivariate Extreme Value Theory (EVT) is the *angular measure*, which provides direct information about the probable ‘directions’ of extremes, that is, the relative contribution of each feature/coordinate of the ‘largest’ observations. Modeling the angular measure in high dimensional problems is a major challenge for the multivariate analysis of rare events. The present chapter proposes a novel methodology aiming at exhibiting a sparsity pattern within the dependence structure of extremes. This is achieved by estimating the amount of mass spread by the angular measure on representative sets of directions, corresponding to specific sub-cones of \mathbb{R}_+^d . This dimension reduction technique paves the way towards scaling up existing multivariate EVT methods. Beyond a non-asymptotic study providing a theoretical validity framework for our method, we propose as a direct application a –first– Anomaly Detection algorithm based on *multivariate* EVT. This algorithm builds a sparse ‘normal profile’ of extreme behaviours, to be confronted with new (possibly abnormal) extreme observations. Illustrative experimental results provide strong empirical evidence of the relevance of our approach. The material of this chapter is based on previous work under review available in Goix et al. (2016c). The empirical part of this work have been published in Goix et al. (2015c).

9.1 Introduction

9.1.1 Context: multivariate extreme values in large dimension

Extreme Value Theory (EVT in abbreviated form) provides a theoretical basis for modeling the tails of probability distributions. In many applied fields where rare events may have a disastrous impact, such as finance, insurance, climate, environmental risk management, network monitoring (Finkenstadt & Rootzén (2003); Smith (2003)) or anomaly detection (Clifton et al. (2011); Lee & Roberts (2008)), the information carried by extremes is crucial. In a multivariate context, the dependence structure of the joint tail is of particular interest, as it gives access *e.g.* to probabilities of a joint excess above high thresholds or to multivariate quantile regions. Also, the

distributional structure of extremes indicates which components of a multivariate quantity may be simultaneously large while the others stay small, which is a valuable piece of information for multi-factor risk assessment or detection of anomalies among other –not abnormal– extreme data.

In a multivariate ‘Peak-Over-Threshold’ setting, realizations of a d -dimensional random vector $\mathbf{Y} = (Y_1, \dots, Y_d)$ are observed and the goal pursued is to learn the conditional distribution of excesses, $[\mathbf{Y} \mid \|\mathbf{Y}\| \geq r]$, above some large threshold $r > 0$. The dependence structure of such excesses is described via the distribution of the ‘directions’ formed by the most extreme observations, the so-called *angular measure*, hereafter denoted by Φ . The latter is defined on the positive orthant of the $d - 1$ dimensional hyper-sphere. To wit, for any region A on the unit sphere (a set of ‘directions’), after suitable standardization of the data (see Section 9.2), $C\Phi(A) \simeq \mathbb{P}(\|\mathbf{Y}\|^{-1}\mathbf{Y} \in A \mid \|\mathbf{Y}\| > r)$, where C is a normalizing constant. Some probability mass may be spread on any sub-sphere of dimension $k < d$, the k -faces of an hyper-cube if we use the infinity norm, which complexifies inference when d is large. To fix ideas, the presence of Φ -mass on a sub-sphere of the type $\{\max_{1 \leq i \leq k} x_i = 1; x_i > 0 (i \leq k); x_{k+1} = \dots = x_d = 0\}$ indicates that the components Y_1, \dots, Y_k may simultaneously be large, while the others are small. An extensive exposition of this multivariate extreme setting may be found *e.g.* in Resnick (1987), Beirlant et al. (2004).

Parametric or semi-parametric modeling and estimation of the structure of multivariate extremes is relatively well documented in the statistical literature, see *e.g.* Coles & Tawn (1991); Fougères et al. (2009); Cooley et al. (2010); Sabourin & Naveau (2012) and the references therein. In a non-parametric setting, there is also an abundant literature concerning consistency and asymptotic normality of estimators of functionals characterizing the extreme dependence structure, *e.g.* extreme value copulas or the *stable tail dependence function* (STDF), see Segers (2012a), Drees & Huang (1998), Embrechts et al. (2000), Einmahl et al. (2012), de Haan & Ferreira (2006). In many applications, it is nevertheless more convenient to work with the angular measure itself, as the latter gives more direct information on the dependence structure and is able to reflect structural simplifying properties (*e.g.* sparsity as detailed below) which would not appear in copulas or in the STDF. However, non-parametric modeling of the angular measure faces major difficulties, stemming from the potentially complexe structure of the latter, especially in a high dimensional setting. Further, from a theoretical point of view, non-parametric estimation of the angular measure has only been studied in the two dimensional case, in Einmahl et al. (2001) and Einmahl & Segers (2009), in an asymptotic framework.

Scaling up multivariate EVT is a major challenge that one faces when confronted to high-dimensional learning tasks, since most multivariate extreme value models have been designed to handle moderate dimensional problems (say, of dimensionality $d \leq 10$). For larger dimensions, simplifying modeling choices are needed, stipulating *e.g.* that only some pre-definite subgroups of components may be concomitantly extremes, or, on the contrary, that all of them must be (see *e.g.* Stephenson (2009) or Sabourin & Naveau (2012)). This curse of dimensionality can be explained, in the context of extreme values analysis, by the relative scarcity of extreme data,

the computational complexity of the estimation procedure and, in the parametric case, by the fact that the dimension of the parameter space usually grows with that of the sample space. This calls for dimensionality reduction devices adapted to multivariate extreme values.

In a wide range of situations, one may expect the occurrence of two phenomena:

- 1- 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 (‘small’ is relative to the total number of groups 2^d).
- 2- 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 .

The main purpose of this chapter is to introduce a data-driven methodology for identifying such faces, so as to reduce the dimensionality of the problem and thus to learn a sparse representation of extreme behaviors. In case hypothesis 2- is not fulfilled, such a sparse ‘profile’ can still be learned, but loses the low dimensional property of its supporting hyper-cubes.

One major issue is that real data generally do not concentrate on sub-spaces of zero Lebesgue measure. This is circumvented by setting to zero any coordinate less than a threshold $\epsilon > 0$, so that the corresponding ‘angle’ is assigned to a lower-dimensional face.

The theoretical results stated in this chapter build on the work of Goix et al. (2015b) exposed in Chapter 8, where non-asymptotic bounds related to the statistical performance of a non-parametric estimator of the STDF, another functional measure of the dependence structure of extremes, are established. However, even in the case of a sparse angular measure, the support of the STDF would not be so, since the latter functional is an integrated version of the former (see (9.6), Section 9.2). Also, in many applications, it is more convenient to work with the angular measure. Indeed, it provides direct information about the probable ‘directions’ of extremes, that is, the relative contribution of each components of the ‘largest’ observations (where ‘large’ may be understood *e.g.* in the sense of the infinity norm on the input space). We emphasize again that estimating these ‘probable relative contributions’ is a major concern in many fields involving the management of risks from multiple sources. To the best of our knowledge, non-parametric estimation of the angular measure has only been treated in the two dimensional case, in Einmahl et al. (2001) and Einmahl & Segers (2009), in an asymptotic framework.

Main contributions. The present contribution extends the non-asymptotic study derived in Chapter 8 to the angular measure of extremes, restricted to a well-chosen representative class of sets, corresponding to lower-dimensional regions of the space. The objective is to learn a representation of the angular measure, rough enough to control the variance in high dimension and accurate enough to gain information about the ‘probable directions’ of extremes. This yields a –first– non-parametric estimate of the angular measure in any dimension, restricted to a class of sub-cones, with a non asymptotic bound on the error. The representation thus obtained is exploited to detect anomalies among extremes.

The proposed algorithm is based on *dimensionality reduction*. We believe that our method can also be used as a preprocessing stage, for dimensionality reduction purpose, before proceeding with a parametric or semi-parametric estimation which could benefit from the structural information issued in the first step. Such applications are beyond the scope of this work and will be the subject of further research.

9.1.2 Application to Anomaly Detection

The framework we develop in this chapter is non-parametric and lies at the intersection of support estimation, density estimation and dimensionality reduction: it consists in learning from training data the support of a distribution, that can be decomposed into sub-cones, hopefully of low dimension each and to which some mass is assigned, according to empirical versions of probability measures on extreme regions.

EVT has been intensively used in anomaly detection in the one-dimensional situation, see for instance Roberts (1999), Roberts (2000), Clifton et al. (2011), Clifton et al. (2008), Lee & Roberts (2008). In the multivariate setup, however, there is –to the best of our knowledge– no anomaly detection method relying on *multivariate* EVT. Until now, the multidimensional case has only been tackled by means of extreme value statistics based on univariate EVT. The major reason is the difficulty to scale up existing multivariate EVT models with the dimensionality. In the present contribution we bridge the gap between the practice of anomaly detection and multivariate EVT by proposing a method which is able to learn a sparse ‘normal profile’ of multivariate extremes and, as such, may be implemented to improve the accuracy of any usual anomaly detection algorithm. Experimental results show that this method significantly improves the performance in extreme regions, as the risk is taken not to uniformly predict as abnormal the most extremal observations, but to learn their dependence structure. These improvements may typically be useful in applications where the cost of false positive errors (*i.e.* false alarms) is very high (*e.g.* predictive maintenance in aeronautics).

The structure of this chapter is as follows. The whys and wherefores of multivariate EVT are explained in the following Section 9.2. A non-parametric estimator of the subfaces’ mass is introduced in Section 9.3, the accuracy of which is investigated by establishing finite sample error bounds relying on VC inequalities tailored to low probability regions. An application to anomaly detection is proposed in Section 9.4, where some background on anomaly detection is provided, followed by a novel anomaly detection algorithm which relies on the above mentioned non-parametric estimator. Experiments on both simulated and real data are performed in Section 9.5. Technical details are deferred at the end of this chapter, Section 9.7.

9.2 Multivariate EVT Framework and Problem Statement

Extreme Value Theory (EVT) develops models to provide a reasonable assessment of the probability of occurrence of rare events. Such models are widely used in fields involving risk management such as Finance, Insurance, Operation Research, Telecommunication or Environmental Sciences for instance. For clarity, we start off with recalling some key notions developed in Chapter 4 pertaining to (multivariate) EVT, that shall be involved in the formulation of the problem next stated and in its subsequent analysis.

First recall the primal assumption of multivariate extreme value theory. For a d -dimensional r.v. $\mathbf{X} = (X^1, \dots, X^d)$ with distribution $\mathbf{F}(\mathbf{x}) := \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d)$, namely $\mathbf{F} \in \mathbf{DA}(\mathbf{G})$ it stipulates the existence of two sequences $\{\mathbf{a}_n, n \geq 1\}$ and $\{\mathbf{b}_n, n \geq 1\}$ in \mathbb{R}^d , the \mathbf{a}_n 's being positive, and a non-degenerate distribution function \mathbf{G} such that

$$\lim_{n \rightarrow \infty} n \mathbb{P} \left(\frac{X^1 - b_n^1}{a_n^1} \geq x_1 \text{ or } \dots \text{ or } \frac{X^d - b_n^d}{a_n^d} \geq x_d \right) = -\log \mathbf{G}(\mathbf{x}) \quad (9.1)$$

for all continuity points $\mathbf{x} \in \mathbb{R}^d$ of \mathbf{G} . Recall also that considering the standardized variables $V^j = 1/(1 - F_j(X^j))$ and $\mathbf{V} = (V^1, \dots, V^d)$, Assumption (9.1) implies the existence of a limit measure μ on $[0, \infty]^d \setminus \{\mathbf{0}\}$ such that

$$n \mathbb{P} \left(\frac{V^1}{n} \geq v_1 \text{ or } \dots \text{ or } \frac{V^d}{n} \geq v_d \right) \xrightarrow[n \rightarrow \infty]{} \mu([\mathbf{0}, \mathbf{v}]^c), \quad (9.2)$$

where $[\mathbf{0}, \mathbf{v}] := [0, v_1] \times \dots \times [0, v_d]$. The dependence structure of the limit \mathbf{G} in (9.1) can then be expressed by means of the so-termed *exponent measure* μ :

$$-\log \mathbf{G}(\mathbf{x}) = \mu \left(\left[\mathbf{0}, \left(\frac{-1}{\log G_1(x_1)}, \dots, \frac{-1}{\log G_d(x_d)} \right) \right]^c \right).$$

The measure μ should be viewed, up to a normalizing factor, as the asymptotic distribution of \mathbf{V} in extreme regions. Also, for any borelian subset A bounded away from $\mathbf{0}$ on which μ is continuous, we have

$$t \mathbb{P}(\mathbf{V} \in tA) \xrightarrow[t \rightarrow \infty]{} \mu(A). \quad (9.3)$$

Using the homogeneity property $\mu(t \cdot) = t^{-1} \mu(\cdot)$, μ can be decomposed into a radial component and an angular component Φ , which are independent from each other. For all $\mathbf{v} = (v_1, \dots, v_d) \in \mathbb{R}^d$, set

$$\begin{cases} R(\mathbf{v}) := \|\mathbf{v}\|_\infty = \max_{i=1}^d v_i, \\ \Theta(\mathbf{v}) := \left(\frac{v_1}{R(\mathbf{v})}, \dots, \frac{v_d}{R(\mathbf{v})} \right) \in S_\infty^{d-1}, \end{cases} \quad (9.4)$$

where S_∞^{d-1} is the positive orthant of the unit sphere in \mathbb{R}^d for the infinity norm. Define the *spectral measure* (also called *angular measure*) by $\Phi(B) = \mu(\{\mathbf{v} : R(\mathbf{v}) > 1, \Theta(\mathbf{v}) \in B\})$. Then, for every $B \subset S_\infty^{d-1}$,

$$\mu\{\mathbf{v} : R(\mathbf{v}) > z, \Theta(\mathbf{v}) \in B\} = z^{-1}\Phi(B). \quad (9.5)$$

In a nutshell, there is a one-to-one correspondence between the exponent measure μ and the angular measure Φ , both of them can be used to characterize the asymptotic tail dependence of the distribution \mathbf{F} (as soon as the margins F_j are known), since

$$\mu([0, \mathbf{x}^{-1}]^c) = \int_{\boldsymbol{\theta} \in S_\infty^{d-1}} \max_j \boldsymbol{\theta}_j x_j \, d\Phi(\boldsymbol{\theta}). \quad (9.6)$$

Recall that here and beyond, operators on vectors are understood component-wise, so that $\mathbf{x}^{-1} = (x_1^{-1}, \dots, x_d^{-1})$. The angular measure can be seen as the asymptotic conditional distribution of the ‘angle’ Θ given that the radius R is large, up to the normalizing constant $\Phi(S_\infty^{d-1})$. Indeed, dropping the dependence on \mathbf{V} for convenience, we have for any *continuity set* A of Φ ,

$$\mathbb{P}(\Theta \in A \mid R > r) = \frac{r\mathbb{P}(\Theta \in A, R > r)}{r\mathbb{P}(R > r)} \xrightarrow{r \rightarrow \infty} \frac{\Phi(A)}{\Phi(S_\infty^{d-1})}. \quad (9.7)$$

9.2.1 Statement of the Statistical Problem

The focus of this work is on the dependence structure in extreme regions of a random vector \mathbf{X} in a multivariate domain of attraction (see (9.1)). This asymptotic dependence is fully described by the exponent measure μ , or equivalently by the spectral measure Φ . The goal of this contribution is to infer a meaningful (possibly sparse) summary of the latter. As shall be seen below, since the support of μ can be naturally partitioned in a specific and interpretable manner, this boils down to accurately recovering the mass spread on each element of the partition. In order to formulate this approach rigorously, additional definitions are required.

Truncated cones. For any non empty subset of features $\alpha \subset \{1, \dots, d\}$, consider the truncated cone (see Fig. 9.1)

$$\mathcal{C}_\alpha = \{\mathbf{v} \geq 0, \|\mathbf{v}\|_\infty \geq 1, v_j > 0 \text{ for } j \in \alpha, v_j = 0 \text{ for } j \notin \alpha\}. \quad (9.8)$$

The corresponding subset of the sphere is

$$\Omega_\alpha = \{\mathbf{x} \in S_\infty^{d-1} : x_i > 0 \text{ for } i \in \alpha, x_i = 0 \text{ for } i \notin \alpha\} = S_\infty^{d-1} \cap \mathcal{C}_\alpha,$$

and we clearly have $\mu(\mathcal{C}_\alpha) = \Phi(\Omega_\alpha)$ for any $\emptyset \neq \alpha \subset \{1, \dots, d\}$. The collection $\{\mathcal{C}_\alpha : \emptyset \neq \alpha \subset \{1, \dots, d\}\}$ forming a partition of the truncated positive orthant $\mathbb{R}_+^d \setminus [0, 1]$, one may

naturally decompose the exponent measure as

$$\mu = \sum_{\emptyset \neq \alpha \subset \{1, \dots, d\}} \mu_\alpha, \quad (9.9)$$

where each component μ_α is concentrated on the untruncated cone corresponding to \mathcal{C}_α . Similarly, the Ω_α 's forming a partition of S_∞^{d-1} , we have

$$\Phi = \sum_{\emptyset \neq \alpha \subset \{1, \dots, d\}} \Phi_\alpha,$$

where Φ_α denotes the restriction of Φ to Ω_α for all $\emptyset \neq \alpha \subset \{1, \dots, d\}$. The fact that mass is spread on \mathcal{C}_α indicates that conditioned upon the event ' $R(\mathbf{V})$ is large' (i.e. an excess of a large radial threshold), the components $V^j (j \in \alpha)$ may be simultaneously large while the other V^j 's ($j \notin \alpha$) are small, with positive probability. Each index subset α thus defines a specific direction in the tail region.

However this interpretation should be handled with care, since for $\alpha \neq \{1, \dots, d\}$, if $\mu(\mathcal{C}_\alpha) > 0$, then \mathcal{C}_α is not a continuity set of μ (it has empty interior), nor Ω_α is a continuity set of Φ . Thus, the quantity $t\mathbb{P}(\mathbf{V} \in t\mathcal{C}_\alpha)$ does not necessarily converge to $\mu(\mathcal{C}_\alpha)$ as $t \rightarrow +\infty$. Actually, if \mathbf{F} is continuous, we have $\mathbb{P}(\mathbf{V} \in t\mathcal{C}_\alpha) = 0$ for any $t > 0$. However, consider for $\epsilon \geq 0$ the ϵ -thickened rectangles

$$R_\alpha^\epsilon = \{\mathbf{v} \geq 0, \|\mathbf{v}\|_\infty \geq 1, v_j > \epsilon \text{ for } j \in \alpha, v_j \leq \epsilon \text{ for } j \notin \alpha\}, \quad (9.10)$$

Since the boundaries of the sets R_α^ϵ are disjoint, only a countable number of them may be discontinuity sets of μ . Hence, the threshold ϵ may be chosen arbitrarily small in such a way that R_α^ϵ is a continuity set of μ . The result stated below shows that nonzero mass on \mathcal{C}_α is the same as nonzero mass on R_α^ϵ for ϵ arbitrarily small.

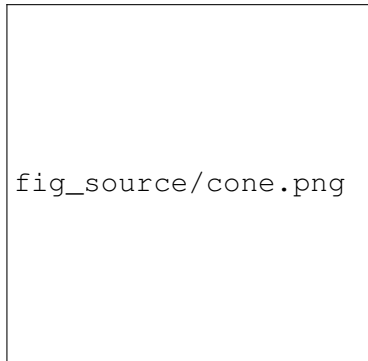


FIGURE 9.1: Truncated cones in 3D

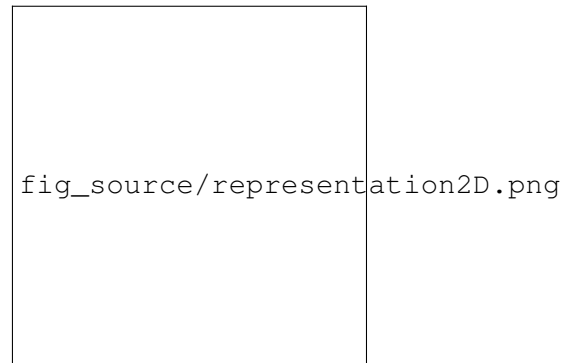


FIGURE 9.2: Truncated ϵ -rectangles in 2D

Lemma 9.1. *For any non empty index subset $\emptyset \neq \alpha \subset \{1, \dots, d\}$, the exponent measure of \mathcal{C}_α is*

$$\mu(\mathcal{C}_\alpha) = \lim_{\epsilon \rightarrow 0} \mu(R_\alpha^\epsilon).$$

Proof. First consider the case $\alpha = \{1, \dots, d\}$. Then R_α^ϵ 's forms an increasing sequence of sets as ϵ decreases and $\mathcal{C}_\alpha = R_\alpha^0 = \bigcup_{\epsilon>0, \epsilon \in \mathbb{Q}} R_\alpha^\epsilon$. The result follows from the ‘continuity from below’ property of the measure μ . Now, for $\epsilon \geq 0$ and $\alpha \subsetneq \{1, \dots, d\}$, consider the sets

$$\begin{aligned} O_\alpha^\epsilon &= \{\mathbf{x} \in \mathbb{R}_+^d : \forall j \in \alpha : x_j > \epsilon\}, \\ N_\alpha^\epsilon &= \{\mathbf{x} \in \mathbb{R}_+^d : \forall j \in \alpha : x_j > \epsilon, \exists j \notin \alpha : x_j > \epsilon\}, \end{aligned}$$

so that $N_\alpha^\epsilon \subset O_\alpha^\epsilon$ and $R_\alpha^\epsilon = O_\alpha^\epsilon \setminus N_\alpha^\epsilon$. Observe also that $\mathcal{C}_\alpha = O_\alpha^0 \setminus N_\alpha^0$. Thus, $\mu(R_\alpha^\epsilon) = \mu(O_\alpha^\epsilon) - \mu(N_\alpha^\epsilon)$, and $\mu(\mathcal{C}_\alpha) = \mu(O_\alpha^0) - \mu(N_\alpha^0)$, so that it is sufficient to show that

$$\mu(N_\alpha^0) = \lim_{\epsilon \rightarrow 0} \mu(N_\alpha^\epsilon), \quad \text{and} \quad \mu(O_\alpha^0) = \lim_{\epsilon \rightarrow 0} \mu(O_\alpha^\epsilon).$$

Notice that the N_α^ϵ 's and the O_α^ϵ 's form two increasing sequences of sets (when ϵ decreases), and that $N_\alpha^0 = \bigcup_{\epsilon>0, \epsilon \in \mathbb{Q}} N_\alpha^\epsilon$, $O_\alpha^0 = \bigcup_{\epsilon>0, \epsilon \in \mathbb{Q}} O_\alpha^\epsilon$. This proves the desired result. \square

We may now make precise the above heuristic interpretation of the quantities $\mu(\mathcal{C}_\alpha)$: the vector $\mathcal{M} = \{\mu(\mathcal{C}_\alpha) : \emptyset \neq \alpha \subset \{1, \dots, d\}\}$ asymptotically describes the dependence structure of the extremal observations. Indeed, by Lemma 9.1, and the discussion above, ϵ may be chosen such that R_α^ϵ is a continuity set of μ , while $\mu(R_\alpha^\epsilon)$ is arbitrarily close to $\mu(\mathcal{C}_\alpha)$. Then, using the characterization (9.3) of μ , the following asymptotic identity holds true:

$$\begin{aligned} \lim_{t \rightarrow \infty} t \mathbb{P}(\|\mathbf{V}\|_\infty \geq t, V^j > \epsilon t \ (j \in \alpha), V^j \leq \epsilon t \ (j \notin \alpha)) &= \mu(R_\alpha^\epsilon) \\ &\simeq \mu(\mathcal{C}_\alpha). \end{aligned} \tag{9.11}$$

Remark 9.2. In terms of conditional probabilities, denoting $R = \|T(\mathbf{X})\|$, where T is the standardization map $\mathbf{X} \mapsto \mathbf{V}$, we have

$$\mathbb{P}(T(\mathbf{X}) \in r R_\alpha^\epsilon \mid R > r) = \frac{r \mathbb{P}(\mathbf{V} \in r R_\alpha^\epsilon)}{r \mathbb{P}(\mathbf{V} \in r([\mathbf{0}, \mathbf{1}]^c))} \xrightarrow{r \rightarrow \infty} \frac{\mu(R_\alpha^\epsilon)}{\mu([\mathbf{0}, \mathbf{1}]^c)},$$

as in (9.7). In other terms,

$$\begin{aligned} \mathbb{P}(V^j > \epsilon r \ (j \in \alpha), V^j \leq \epsilon r \ (j \notin \alpha) \mid \|\mathbf{V}\|_\infty \geq r) &\xrightarrow{r \rightarrow \infty} C \mu(R_\alpha^\epsilon) \\ &\simeq C \mu(\mathcal{C}_\alpha), \end{aligned} \tag{9.12}$$

where $C = 1/\Phi(S_\infty^{d-1}) = 1/\mu([\mathbf{0}, \mathbf{1}]^c)$. This clarifies the meaning of ‘large’ and ‘small’ in the heuristic explanation given above.

Problem statement. As explained above, our goal is to describe the dependence on extreme regions by investigating the structure of μ (or, equivalently, that of Φ). More precisely, the aim is twofold. First, recover a rough approximation of the support of Φ based on the partition $\{\Omega_\alpha, \alpha \subset \{1, \dots, d\}, \alpha \neq \emptyset\}$, that is, determine which Ω_α 's have nonzero mass, or equivalently, which μ'_α 's (resp. Φ_α 's) are nonzero. This support estimation is potentially sparse (if a small

number of Ω_α have non-zero mass) and possibly low-dimensional (if the dimension of the sub-cones Ω_α with non-zero mass is low). The second objective is to investigate how the exponent measure μ spreads its mass on the \mathcal{C}_α 's, the theoretical quantity $\mu(\mathcal{C}_\alpha)$ indicating to which extent extreme observations may occur in the 'direction' α for $\emptyset \neq \alpha \subset \{1, \dots, d\}$. These two goals are achieved using empirical versions of the angular measure defined in Section 9.3.1, evaluated on the ϵ -thickened rectangles R_α^ϵ . Formally, we wish to recover the $(2^d - 1)$ -dimensional unknown vector

$$\mathcal{M} = \{\mu(\mathcal{C}_\alpha) : \emptyset \neq \alpha \subset \{1, \dots, d\}\} \quad (9.13)$$

from $\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{i.i.d.}{\sim} \mathbf{F}$ and build an estimator $\widehat{\mathcal{M}}$ such that

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

is small with large probability. In view of Lemma 9.1, (biased) estimates of \mathcal{M} 's components are built from an empirical version of the exponent measure, evaluated on the ϵ -thickened rectangles R_α^ϵ (see Section 9.3.1 below). As a by-product, one obtains an estimate of the support of the limit measure μ ,

$$\bigcup_{\alpha: \widehat{\mathcal{M}}(\alpha) > 0} \mathcal{C}_\alpha.$$

The results stated in the next section are non-asymptotic and sharp bounds are given by means of VC inequalities tailored to low probability regions.

9.2.2 Regularity Assumptions

Beyond the existence of the limit measure μ (*i.e.* multivariate regular variation of \mathbf{V} 's distribution, see (9.2)), and thus, existence of an angular measure Φ (see (9.5)), three additional assumptions are made, which are natural when estimation of the support of a distribution is considered.

Assumption 1. The margins of \mathbf{X} have continuous c.d.f., namely F_j , $1 \leq j \leq d$ is continuous.

Assumption 1 is widely used in the context of non-parametric estimation of the dependence structure (see *e.g.* Einmahl & Segers (2009)): it ensures that the transformed variables $V^j = (1 - F_j(X^j))^{-1}$ (*resp.* $U^j = 1 - F_j(X^j)$) have indeed a standard Pareto distribution, $\mathbb{P}(V^j > x) = 1/x$, $x \geq 1$ (*resp.* the U^j 's are uniform variables).

For any non empty subset α of $\{1, \dots, d\}$, one denotes by dx_α the Lebesgue measure on \mathcal{C}_α and write $dx_\alpha = dx_{i_1} \dots dx_{i_k}$, when $\alpha = \{i_1, \dots, i_k\}$. For convenience, we also write $dx_{\alpha \setminus i}$ instead of $dx_{\alpha \setminus \{i\}}$.

Assumption 2. Each component μ_α of (9.9) is absolutely continuous w.r.t. Lebesgue measure dx_α on \mathcal{C}_α .

Assumption 2 has a very convenient consequence regarding Φ : the fact that the exponent measure μ spreads no mass on subsets of the form $\{\mathbf{x} : \|\mathbf{x}\|_\infty \geq 1, x_{i_1} = \dots = x_{i_r} \neq 0\}$ with $r \geq 2$, implies that the spectral measure Φ spreads no mass on edges $\{\mathbf{x} : \|\mathbf{x}\|_\infty = 1, x_{i_1} = \dots = x_{i_r} = 1\}$ with $r \geq 2$. This is summarized by the following result.

Lemma 9.3. *Under Assumption 2, the following assertions holds true.*

- Φ is concentrated on the (disjoint) edges

$$\Omega_{\alpha, i_0} = \{\mathbf{x} : \|\mathbf{x}\|_\infty = 1, x_{i_0} = 1, 0 < x_i < 1 \text{ for } i \in \alpha \setminus \{i_0\} \\ x_i = 0 \text{ for } i \notin \alpha\} \quad (9.14)$$

for $i_0 \in \alpha, \emptyset \neq \alpha \subset \{1, \dots, d\}$.

- The restriction Φ_{α, i_0} of Φ to Ω_{α, i_0} is absolutely continuous w.r.t. the Lebesgue measure $dx_{\alpha \setminus i_0}$ on the cube's edges, whenever $|\alpha| \geq 2$.

Proof. The first assertion straightforwardly results from the discussion above. Turning to the second point, consider any measurable set $D \subset \Omega_{\alpha, i_0}$ such that $\int_D dx_{\alpha \setminus i_0} = 0$. Then the induced truncated cone $\tilde{D} = \{\mathbf{v} : \|\mathbf{v}\|_\infty \geq 1, \mathbf{v}/\|\mathbf{v}\|_\infty \in D\}$ satisfies $\int_{\tilde{D}} dx_\alpha = 0$ and belongs to \mathcal{C}_α . Thus, by virtue of Assumption 2, $\Phi_{\alpha, i_0}(D) = \Phi_\alpha(D) = \mu_\alpha(\tilde{D}) = 0$. \square

It follows from Lemma 9.3 that the angular measure Φ decomposes as $\Phi = \sum_\alpha \sum_{i_0 \in \alpha} \Phi_{\alpha, i_0}$ and that there exist densities $\frac{d\Phi_{\alpha, i_0}}{dx_{\alpha \setminus i_0}}, |\alpha| \geq 2, i_0 \in \alpha$, such that for all $B \subset \Omega_\alpha, |\alpha| \geq 2$,

$$\Phi(B) = \Phi_\alpha(B) = \sum_{i_0 \in \alpha} \int_{B \cap \Omega_{\alpha, i_0}} \frac{d\Phi_{\alpha, i_0}}{dx_{\alpha \setminus i_0}}(x) dx_{\alpha \setminus i_0}. \quad (9.15)$$

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

$$M_\beta = \sup_{i \in \beta} \sup_{x \in \Omega_{\beta, i}} \frac{d\Phi_{\beta, i}}{dx_{\beta \setminus i}}(x). \quad (9.16)$$

Assumption 3. (SPARSE SUPPORT) The angular density is uniformly bounded on S_∞^{d-1} ($\forall |\beta| \geq 2, M_\beta < \infty$), and there exists a constant $M > 0$, such that we have $\sum_{|\beta| \geq 2} M_\beta < M$, where the sum is over subsets β of $\{1, \dots, d\}$ which contain at least two elements.

Remark 9.4. The constant M is problem dependent. However, in the case where our representation \mathcal{M} defined in (9.13) is the most informative about the angular measure, that is, when the density of Φ_α is constant on Ω_α , we have $M \leq d$: Indeed, in such a case, $M \leq \sum_{|\beta| \geq 2} M_\beta |\beta| = \sum_{|\beta| \geq 2} \Phi(\Omega_\beta) \leq \sum_\beta \Phi(\Omega_\beta) \leq \mu([0, 1]^c)$. The equality inside the last expression comes from

the fact that the Lebesgue measure of a sub-sphere Ω_α is $|\alpha|$, for $|\alpha| \geq 2$. Indeed, using the notations defined in Lemma 9.3, $\Omega_\alpha = \bigsqcup_{i_0 \in \alpha} \Omega_{\alpha, i_0}$, each of the edges Ω_{α, i_0} being unit hypercube. Now, $\mu([0, 1]^c) \leq \mu(\{v, \exists j, v_j > 1\}) \leq d\mu(\{v, v_1 > 1\}) \leq d$.

Note that the summation $\sum_{|\beta| \geq 2} M_\beta |\beta|$ is smaller than d despite the (potentially large) factors $|\beta|$. Considering $\sum_{|\beta| \geq 2} M_\beta$ is thus reasonable: in particular, M will be small when only few Ω_α 's have non-zero Φ -mass, namely when the representation vector \mathcal{M} defined in (9.13) is sparse.

Assumption 3 is naturally involved in the derivation of upper bounds on the error made when approximating $\mu(\mathcal{C}_\alpha)$ by the empirical counterpart of $\mu(R_\alpha^\epsilon)$. The estimation error bound derived in Section 9.3 depends on the sparsity constant M .

9.3 A non-parametric estimator of the subcones' mass : definition and preliminary results

In this section, an estimator $\widehat{\mathcal{M}}(\alpha)$ of each of the sub-cones' mass $\mu(\mathcal{C}_\alpha)$, $\emptyset \neq \alpha \subset \{1, \dots, d\}$, is proposed, based on observations $\mathbf{X}_1, \dots, \mathbf{X}_n$, *i.i.d.* copies of $\mathbf{X} \sim \mathbf{F}$. Bounds on the error $\|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty$ are established. In the remaining of this chapter, we work under Assumption 1 (continuous margins, see Section 9.2.2). Assumptions 2 and 3 are not necessary to prove a preliminary result on a class of rectangles (Proposition 9.8 and Corollary 9.9). However, they are required to bound the bias induced by the tolerance parameter ϵ (in Lemma 9.10, Proposition 9.11 and in the main result, Theorem 9.12).

9.3.1 A natural empirical version of the exponent measure μ

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

$$\widehat{\mathbf{V}}_i = \left((1 - \widehat{F}_j(X_i^j))^{-1} \right)_{1 \leq j \leq d},$$

where $\widehat{F}_j(x) = (1/n) \sum_{i=1}^n \mathbf{1}_{\{X_i^j < x\}}$. We denote by T (*resp.* \widehat{T}) the standardization (*resp.* the empirical standardization),

$$T(\mathbf{x}) = \left(\frac{1}{1 - F_j(x^j)} \right)_{1 \leq j \leq d} \quad \text{and} \quad \widehat{T}(\mathbf{x}) = \left(\frac{1}{1 - \widehat{F}_j(x^j)} \right)_{1 \leq j \leq d}. \quad (9.17)$$

The empirical probability distribution of the rank-transformed data is then given by

$$\widehat{\mathbb{P}}_n = (1/n) \sum_{i=1}^n \delta_{\widehat{\mathbf{V}}_i}.$$

Since for a μ -continuity set A bounded away from 0, $t \mathbb{P}(\mathbf{V} \in tA) \rightarrow \mu(A)$ as $t \rightarrow \infty$, see (9.3), a natural empirical version of μ is defined as

$$\mu_n(A) = \frac{n}{k} \widehat{\mathbb{P}}_n\left(\frac{n}{k}A\right) = \frac{1}{k} \sum_{i=1}^n \mathbf{1}_{\{\widehat{\mathbf{V}}_i \in \frac{n}{k}A\}}. \quad (9.18)$$

Here and throughout, we place ourselves in the asymptotic setting stipulating that $k = k(n) > 0$ is such that $k \rightarrow \infty$ and $k = o(n)$ as $n \rightarrow \infty$. The ratio n/k plays the role of a large radial threshold. Note that this estimator is commonly used in the field of non-parametric estimation of the dependence structure, see *e.g.* Einmahl & Segers (2009).

9.3.2 Accounting for the non asymptotic nature of data: epsilon-thickening.

Since the cones \mathcal{C}_α have zero Lebesgue measure, and since, under Assumption 1, the margins are continuous, the cones are not likely to receive any empirical mass, so that simply counting points in $\frac{n}{k}\mathcal{C}_\alpha$ is not an option: with probability one, only the largest dimensional cone (the central one, corresponding to $\alpha = \{1, \dots, d\}$) will be hit. In view of Subsection 9.2.1 and Lemma 9.1, it is natural to introduce a tolerance parameter $\epsilon > 0$ and to approximate the asymptotic mass of \mathcal{C}_α with the non-asymptotic mass of R_α^ϵ . We thus define the non-parametric estimator $\widehat{\mathcal{M}}(\alpha)$ of $\mu(\mathcal{C}_\alpha)$ as


$$\widehat{\mathcal{M}}(\alpha) = \mu_n(R_\alpha^\epsilon), \quad \emptyset \neq \alpha \subset \{1, \dots, d\}. \quad (9.19)$$

Evaluating $\widehat{\mathcal{M}}(\alpha)$ boils down (see (9.18)) to counting points in $(n/k) R_\alpha^\epsilon$, as illustrated in Figure 9.3. The estimate $\widehat{\mathcal{M}}(\alpha)$ is thus a (voluntarily ϵ -biased) natural estimator of $\Phi(\Omega_\alpha) = \mu(\mathcal{C}_\alpha)$.

The coefficients $(\widehat{\mathcal{M}}(\alpha))_{\alpha \subset \{1, \dots, d\}}$ related to the cones \mathcal{C}_α constitute a summary representation of the dependence structure. This representation is sparse as soon as the $\mu_n^{\alpha, \epsilon}$ are positive only for a few groups of features α (compared to the total number of groups or sub-cones, 2^d namely). It is low-dimensional as soon as each of these groups α is of small cardinality, or equivalently the corresponding sub-cones are low-dimensional compared with d .

In fact, $\widehat{\mathcal{M}}(\alpha)$ is (up to a normalizing constant) an empirical version of the conditional probability that $T(\mathbf{X})$ belongs to the rectangle rR_α^ϵ , given that $\|T(\mathbf{X})\|$ exceeds a large threshold r . Indeed, as explained in Remark 9.2,

$$\mathcal{M}(\alpha) = \lim_{r \rightarrow \infty} \mu([\mathbf{0}, \mathbf{1}]^c) \mathbb{P}(T(\mathbf{X}) \in rR_\alpha^\epsilon \mid \|T(\mathbf{X})\| \geq r). \quad (9.20)$$



fig_source/representation2D_nk_rect.png

FIGURE 9.3: Estimation procedure

The remaining of this section is devoted to obtaining non-asymptotic upper bounds on the error $\|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty$. The main result is stated in Theorem 9.12. Before all, notice that the error may be obviously decomposed as the sum of a stochastic term and a bias term inherent to the ϵ -thickening approach:

$$\begin{aligned} \|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty &= \max_{\alpha} |\mu_n(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)| \\ &\leq \max_{\alpha} |\mu - \mu_n|(R_\alpha^\epsilon) + \max_{\alpha} |\mu(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)|. \end{aligned} \quad (9.21)$$

Here and beyond, for notational convenience, we simply denotes ‘ α ’ for ‘ α non empty subset of $\{1, \dots, d\}$ ’. The main steps of the argument leading to Theorem 9.12 are as follows. First, obtain a uniform upper bound on the error $|\mu_n - \mu|$ restricted to a well chosen VC class of rectangles (Subsection 9.3.3), and deduce an uniform bound on $|\mu_n - \mu|(R_\alpha^\epsilon)$ (Subsection 9.3.4). Finally, using the regularity assumptions (Assumption 2 and Assumption 3), bound the difference $|\mu(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)|$ (Subsection 9.3.5).

9.3.3 Preliminaries: uniform approximation over a VC-class of rectangles

This subsection builds on the theory developed in Chapter 8, where a non-asymptotic bound is stated on the estimation of the stable tail dependence function (defined in (4.12)). The STDFI is related to the class of sets of the form $[0, \mathbf{v}]^c$ (or $[\mathbf{u}, \infty]^c$ depending on which standardization is

used), and an equivalent definition is

$$l(\mathbf{x}) := \lim_{t \rightarrow \infty} t \tilde{F}(t^{-1} \mathbf{x}) = \mu([\mathbf{0}, \mathbf{x}^{-1}]^c) \quad (9.22)$$

with $\tilde{F}(\mathbf{x}) = (1 - F)((1 - F_1)^\leftarrow(x_1), \dots, (1 - F_d)^\leftarrow(x_d))$. Here the notation $(1 - F_j)^\leftarrow(x_j)$ denotes the quantity $\sup\{y : 1 - F_j(y) \geq x_j\}$. Recall that the marginally uniform variable \mathbf{U} is defined by $U^j = 1 - F_j(X^j)$ ($1 \leq j \leq d$). Then in terms of standardized variables U^j ,

$$\tilde{F}(\mathbf{x}) = \mathbb{P}\left(\bigcup_{j=1}^d \{U^j < x_j\}\right) = \mathbb{P}(\mathbf{U} \in [\mathbf{x}, \infty[^c) = \mathbb{P}(\mathbf{V} \in [\mathbf{0}, \mathbf{x}^{-1}]^c). \quad (9.23)$$

A natural estimator of l is its empirical version defined as follows, see Huang (1992), Qi (1997), Drees & Huang (1998), Einmahl et al. (2006), Goix et al. (2015b):

$$l_n(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{X_i^1 \geq X_{(n - \lfloor kx_1 \rfloor + 1)}^1 \text{ or } \dots \text{ or } X_i^d \geq X_{(n - \lfloor kx_d \rfloor + 1)}^d\}}. \quad (9.24)$$

The expression is indeed suggested by the definition of l in (9.22), with all distribution functions and univariate quantiles replaced by their empirical counterparts, and with t replaced by n/k . The following lemma allows to derive alternative expressions for the empirical version of the STDF.

Lemma 9.5. *Consider the rank transformed variables $\hat{\mathbf{U}}_i = (\hat{\mathbf{V}}_i)^{-1} = (1 - \hat{F}_j(X_i^j))_{1 \leq j \leq d}$ for $i = 1, \dots, n$. Then, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, d\}$, with probability one,*

$$\hat{U}_i^j \leq \frac{k}{n} x_j^{-1} \Leftrightarrow \hat{V}_i^j \geq \frac{n}{k} x_j \Leftrightarrow X_i^j \geq X_{(n - \lfloor kx_j^{-1} \rfloor + 1)}^j \Leftrightarrow U_i^j \leq U_{(\lfloor kx_j^{-1} \rfloor)}^j.$$

The proof of Lemma 9.5 is standard and is provided in 9.7 for completeness. By Lemma 9.5, the following alternative expression of $l_n(\mathbf{x})$ holds true:

$$l_n(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{U_i^1 \leq U_{(\lfloor kx_1 \rfloor)}^1 \text{ or } \dots \text{ or } U_i^d \leq U_{(\lfloor kx_d \rfloor)}^d\}} = \mu_n([\mathbf{0}, \mathbf{x}^{-1}]^c). \quad (9.25)$$

Thus, bounding the error $|\mu_n - \mu|([\mathbf{0}, \mathbf{x}^{-1}]^c)$ is the same as bounding $|l_n - l|(\mathbf{x})$.

Asymptotic properties of this empirical counterpart have been studied in Huang (1992), Drees & Huang (1998), Embrechts et al. (2000) and de Haan & Ferreira (2006) in the bivariate case, and Qi (1997), Einmahl et al. (2012). in the general multivariate case. In Goix et al. (2015b), a non-asymptotic bound is established on the maximal deviation

$$\sup_{0 \leq \mathbf{x} \leq T} |l(\mathbf{x}) - l_n(\mathbf{x})|$$

for a fixed $T > 0$, or equivalently on

$$\sup_{1/T \leq \mathbf{x}} |\mu([\mathbf{0}, \mathbf{x}]^c) - \mu_n([\mathbf{0}, \mathbf{x}]^c)|.$$

The exponent measure μ is indeed easier to deal with when restricted to the class of sets of the form $[\mathbf{0}, \mathbf{x}]^c$, which is fairly simple in the sense that it has finite VC dimension.

In the present work, an important step is to bound the error on the class of ϵ -thickened rectangles R_α^ϵ . This is achieved by using a more general class $R(\mathbf{x}, \mathbf{z}, \alpha, \beta)$, which includes (contrary to the collection of sets $[\mathbf{0}, \mathbf{x}]^c$) the R_α^ϵ 's. This flexible class is defined by

$$R(\mathbf{x}, \mathbf{z}, \alpha, \beta) = \left\{ \mathbf{y} \in [0, \infty]^d, \begin{array}{l} y_j \geq x_j \text{ for } j \in \alpha, \\ y_j < z_j \text{ for } j \in \beta \end{array} \right\}, \quad \mathbf{x}, \mathbf{z} \in [0, \infty]^d. \quad (9.26)$$

Thus,

$$\mu_n(R(\mathbf{x}, \mathbf{z}, \alpha, \beta)) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{\widehat{V}_i^j \geq \frac{n}{k} x_j \text{ for } j \in \alpha \text{ and } \widehat{V}_i^j < \frac{n}{k} x_j \text{ for } j \in \beta\}}.$$

Then, define the functional $g_{\alpha, \beta}$ (which plays the same role as the STDF) as follows: for $\mathbf{x} \in [0, \infty]^d \setminus \{\infty\}$, $\mathbf{z} \in [0, \infty]^d$, $\alpha \subset \{1, \dots, d\} \setminus \emptyset$ and $\beta \subset \{1, \dots, d\}$, let

$$g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) = \lim_{t \rightarrow \infty} t \tilde{F}_{\alpha, \beta}(t^{-1} \mathbf{x}, t^{-1} \mathbf{z}), \quad \text{with} \quad (9.27)$$

$$\tilde{F}_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) = \mathbb{P} \left[\{U^j \leq x_j \text{ for } j \in \alpha\} \cap \{U^j > z_j \text{ for } j \in \beta\} \right]. \quad (9.28)$$

Notice that $\tilde{F}_{\alpha, \beta}(\mathbf{x}, \mathbf{z})$ is an extension of the non-asymptotic approximation \tilde{F} in (9.22). By (9.27) and (9.28), we have

$$\begin{aligned} g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) &= \lim_{t \rightarrow \infty} t \mathbb{P} \left[\{U^j \leq t^{-1} x_j \text{ for } j \in \alpha\} \cap \{U^j > t^{-1} z_j \text{ for } j \in \beta\} \right] \\ &= \lim_{t \rightarrow \infty} t \mathbb{P} [\mathbf{V} \in tR(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta)], \end{aligned}$$

so that using (9.3),

$$g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) = \mu([R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta)]). \quad (9.29)$$

The following lemma makes the relation between $g_{\alpha, \beta}$ and the angular measure Φ explicit. Its proof is given in 9.7.

Lemma 9.6. *The function $g_{\alpha, \beta}$ can be represented as follows:*

$$g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) = \int_{S^{d-1}} \left(\bigwedge_{j \in \alpha} w_j x_j - \bigvee_{j \in \beta} w_j z_j \right)_+ \Phi(d\mathbf{w}),$$

where $u \wedge v = \min\{u, v\}$, $u \vee v = \max\{u, v\}$ and $u_+ = \max\{u, 0\}$ for any $(u, v) \in \mathbb{R}^2$. Thus, $g_{\alpha, \beta}$ is homogeneous and satisfies

$$|g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) - g_{\alpha, \beta}(\mathbf{x}', \mathbf{z}')| \leq \sum_{j \in \alpha} |x_j - x'_j| + \sum_{j \in \beta} |z_j - z'_j|,$$

Remark 9.7. Lemma 9.6 shows that the functional $g_{\alpha, \beta}$, which plays the same role as the STDF, enjoys a Lipschitz property.

We now define the empirical counterpart of $g_{\alpha, \beta}$ (mimicking that of the empirical STDF l_n in (9.24)) by

$$g_{n, \alpha, \beta}(\mathbf{x}, \mathbf{z}) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{X_i^j \geq X_{(n - \lfloor kx_j \rfloor + 1)}^j \text{ for } j \in \alpha \text{ and } X_i^j < X_{(n - \lfloor kx_j \rfloor + 1)}^j \text{ for } j \in \beta\}}. \quad (9.30)$$

As it is the case for the empirical STDF (see (9.25)), $g_{n, \alpha, \beta}$ has an alternative expression

$$\begin{aligned} g_{n, \alpha, \beta}(\mathbf{x}, \mathbf{z}) &= \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{U_i^j \leq U_{(\lfloor kx_j \rfloor)}^j \text{ for } j \in \alpha \text{ and } U_i^j > U_{(\lfloor kx_j \rfloor)}^j \text{ for } j \in \beta\}} \\ &= \mu_n(R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta)), \end{aligned} \quad (9.31)$$

where the last equality comes from the equivalence $\widehat{V}_i^j \geq \frac{n}{k} x_j \Leftrightarrow U_i^j \leq U_{(\lfloor kx_j \rfloor)}^j$ (Lemma 9.5) and from the expression $\mu_n(\cdot) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\widehat{\mathbf{V}}_i \in \frac{n}{k}(\cdot)}$, definition (9.18).

The proposition below extends the result of Goix et al. (2015b), by deriving an analogue upper bound on the maximal deviation

$$\max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} |g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) - g_{n, \alpha, \beta}(\mathbf{x}, \mathbf{z})|,$$

or equivalently on

$$\max_{\alpha, \beta} \sup_{1/T \leq \mathbf{x}, \mathbf{z}} |\mu(R(\mathbf{x}, \mathbf{z}, \alpha, \beta)) - \mu_n(R(\mathbf{x}, \mathbf{z}, \alpha, \beta))|.$$

Here and beyond we simply denote ‘ α, β ’ for ‘ α non-empty subset of $\{1, \dots, d\} \setminus \emptyset$ and β subset of $\{1, \dots, d\}$ ’. We also recall that comparison operators between two vectors (or between a vector and a real number) are understood component-wise, *i.e.* ‘ $\mathbf{x} \leq \mathbf{z}$ ’ means ‘ $x_j \leq z_j$ for all $1 \leq j \leq d$ ’ and for any real number T , ‘ $\mathbf{x} \leq T$ ’ means ‘ $x_j \leq T$ for all $1 \leq j \leq d$ ’.

Proposition 9.8. *Let $T \geq \frac{7}{2}(\frac{\log d}{k} + 1)$, and $\delta \geq e^{-k}$. Then there is a universal constant C , such that for each $n > 0$, with probability at least $1 - \delta$,*

$$\begin{aligned} \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} |g_{n, \alpha, \beta}(\mathbf{x}, \mathbf{z}) - g_{\alpha, \beta}(\mathbf{x}, \mathbf{z})| &\leq Cd \sqrt{\frac{2T}{k} \log \frac{d+3}{\delta}} \\ &+ \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq 2T} \left| \frac{n}{k} \tilde{F}_{\alpha, \beta}(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z}) - g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) \right|. \end{aligned} \quad (9.32)$$

The second term on the right hand side of the inequality is an asymptotic bias term which goes to 0 as $n \rightarrow \infty$ (see Remark 9.24).

The proof follows the same lines as that of Theorem 6 in Goix et al. (2015b) and is detailed in 9.7. Here is the main argument.

The empirical estimator is based on the empirical measure of ‘extreme’ regions, which are hit only with low probability. It is thus enough to bound maximal deviations on such low probability regions. The key consists in choosing an adaptive VC class which only covers the latter regions (after standardization to uniform margins), namely a VC class composed of sets of the kind $\frac{k}{n}R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta)^{-1}$. In Goix et al. (2015b), VC-type inequalities have been established that incorporate p , the probability of hitting the class at all. Applying these inequalities to the particular class of rectangles gives the result.

9.3.4 Bounding empirical deviations over thickened rectangles

The aim of this subsection is to bound $|\mu_n - \mu|(R_\alpha^\epsilon)$ uniformly over α exploiting the previously established bound on the deviations on rectangles, to obtain another uniform bound for $|\mu_n - \mu|(R_\alpha^\epsilon)$, for $\epsilon > 0$ and $\alpha \subset \{1, \dots, d\}$. In the remainder of the chapter, $\bar{\alpha}$ denotes the complementary set of α in $\{1, \dots, d\}$. Notice that directly from their definitions (9.10) and (9.26), R_α^ϵ and $R(\mathbf{x}, \mathbf{z}, \alpha, \beta)$ are linked by:

$$R_\alpha^\epsilon = R(\epsilon, \epsilon, \alpha, \bar{\alpha}) \cap [0, 1]^c = R(\epsilon, \epsilon, \alpha, \bar{\alpha}) \setminus R(\epsilon, \tilde{\epsilon}, \alpha, \{1, \dots, d\})$$

where $\tilde{\epsilon}$ is defined by $\tilde{\epsilon}_j = \mathbb{1}_{j \in \alpha} + \epsilon \mathbb{1}_{j \notin \alpha}$ for all $j \in \{1, \dots, d\}$. Indeed, we have: $R(\epsilon, \epsilon, \alpha, \bar{\alpha}) \cap [0, 1] = R(\epsilon, \tilde{\epsilon}, \alpha, \{1, \dots, d\})$. As a result, for $\epsilon < 1$,

$$\sup_{\epsilon \leq \mathbf{x}, \mathbf{z}} |\mu_n - \mu|(R_\alpha^\epsilon) \leq 2 \sup_{\epsilon \leq \mathbf{x}, \mathbf{z}} |\mu_n - \mu|(R(\mathbf{x}, \mathbf{z}, \alpha, \bar{\alpha})).$$

On the other hand, from (9.31) and (9.29) we have

$$\sup_{\epsilon \leq \mathbf{x}, \mathbf{z}} |\mu_n - \mu|(R(\mathbf{x}, \mathbf{z}, \alpha, \bar{\alpha})) = \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq \epsilon^{-1}} |g_{n, \alpha, \bar{\alpha}}(\mathbf{x}, \mathbf{z}) - g_{\alpha, \bar{\alpha}}(\mathbf{x}, \mathbf{z})|.$$

Then Proposition 9.8 applies with $T = 1/\epsilon$ and the following result holds true.

Corollary 9.9. *Let $0 < \epsilon \leq (\frac{7}{2}(\frac{\log d}{k} + 1))^{-1}$, and $\delta \geq e^{-k}$. Then there is a universal constant C , such that for each $n > 0$, with probability at least $1 - \delta$,*

$$\begin{aligned} \max_{\alpha} \sup_{\epsilon \leq \mathbf{x}, \mathbf{z}} |(\mu_n - \mu)(R_\alpha^\epsilon)| &\leq Cd \sqrt{\frac{1}{\epsilon k} \log \frac{d+3}{\delta}} \\ &+ \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq 2\epsilon^{-1}} \left| \frac{n}{k} \tilde{F}_{\alpha, \beta}(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z}) - g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) \right|. \end{aligned} \quad (9.33)$$

9.3.5 Bounding the bias induced by thickened rectangles

In this section, the aim is to bound $|\mu(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)|$ uniformly over α ; in other words, to derive an upper bound on the bias induced by handling ϵ -thickened rectangles. As the rectangles R_α^ϵ defined in (9.10) do not correspond to any set of angles on the sphere S_∞^{d-1} , we also define the (ϵ, ϵ') -thickened cones

$$\mathcal{C}_\alpha^{\epsilon, \epsilon'} = \{\mathbf{v} \geq 0, \|\mathbf{v}\|_\infty \geq 1, v_j > \epsilon \|\mathbf{v}\|_\infty \text{ for } j \in \alpha, v_j \leq \epsilon' \|\mathbf{v}\|_\infty \text{ for } j \notin \alpha\}, \quad (9.34)$$

which verify $\mathcal{C}_\alpha^{\epsilon, 0} \subset R_\alpha^\epsilon \subset \mathcal{C}_\alpha^{0, \epsilon}$. Define the corresponding (ϵ, ϵ') -thickened sub-sphere

$$\Omega_\alpha^{\epsilon, \epsilon'} = \{\mathbf{x} \in S_\infty^{d-1}, x_i > \epsilon \text{ for } i \in \alpha, x_i \leq \epsilon' \text{ for } i \notin \alpha\} = \mathcal{C}_\alpha^{\epsilon, \epsilon'} \cap S_\infty^{d-1}. \quad (9.35)$$

It is then possible to approximate rectangles R_α^ϵ by the cones $\mathcal{C}_\alpha^{\epsilon, 0}$ and $\mathcal{C}_\alpha^{0, \epsilon}$, and then $\mu(R_\alpha^\epsilon)$ by $\Phi(\Omega_\alpha^{\epsilon, \epsilon'})$ in the sense that

$$\Phi(\Omega_\alpha^{\epsilon, 0}) = \mu(\mathcal{C}_\alpha^{\epsilon, 0}) \leq \mu(R_\alpha^\epsilon) \leq \mu(\mathcal{C}_\alpha^{0, \epsilon}) = \Phi(\Omega_\alpha^{0, \epsilon}). \quad (9.36)$$

The next result (proved in 9.7) is a preliminary step toward a bound on $|\mu(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)|$. It is easier to use the absolute continuity of Φ instead of that of μ , since the rectangles R_α^ϵ are not bounded contrary to the sub-spheres $\Omega_\alpha^{\epsilon, \epsilon'}$.

Lemma 9.10. *For every $\emptyset \neq \alpha \subset \{1, \dots, d\}$ and $0 < \epsilon, \epsilon' < 1/2$, we have*

$$|\Phi(\Omega_\alpha^{\epsilon, \epsilon'}) - \Phi(\Omega_\alpha)| \leq M|\alpha|^2\epsilon + Md\epsilon'.$$

Now, notice that

$$\Phi(\Omega_\alpha^{\epsilon, 0}) - \Phi(\Omega_\alpha) \leq \mu(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha) \leq \Phi(\Omega_\alpha^{0, \epsilon}) - \Phi(\Omega_\alpha).$$

We obtain the following proposition.

Proposition 9.11. *For every non empty set of indices $\emptyset \neq \alpha \subset \{1, \dots, d\}$ and $\epsilon > 0$,*

$$|\mu(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)| \leq Md^2\epsilon$$

9.3.6 Main result

We can now state the main result of the contribution, revealing the accuracy of the estimate (9.19).

Theorem 9.12. *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 $\epsilon \leq (\frac{7}{2}(\frac{\log d}{k} + 1))^{-1}$, the following inequality holds true with*

probability greater than $1 - \delta$:

$$\begin{aligned} \|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty \leq & Cd \left(\sqrt{\frac{1}{\epsilon k} \log \frac{d}{\delta}} + Md\epsilon \right) \\ & + 4 \max_{\substack{\alpha \subset \{1, \dots, d\} \\ \alpha \neq \emptyset}} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq \frac{2}{\epsilon}} \left| \frac{n}{k} \tilde{F}_{\alpha, \bar{\alpha}}\left(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z}\right) - g_{\alpha, \bar{\alpha}}(\mathbf{x}, \mathbf{z}) \right|. \end{aligned}$$

Note that $\frac{7}{2}(\frac{\log d}{k} + 1)$ is smaller than 4 as soon as $\log d/k < 1/7$, so that a sufficient condition on ϵ is $\epsilon < 1/4$. The last term in the right hand side is a bias term which goes to zero as $n \rightarrow \infty$ (see Remark 9.24). The term $Md\epsilon$ is also a bias term, which represents the bias induced by considering ϵ -thickened rectangles. It depends linearly on the sparsity constant M defined in Assumption 3. 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. Considering classical inequalities in empirical process theory such as VC-bounds, it is thus no surprise to obtain one in $O(1/\sqrt{k})$. Too large values of k tend to yield a large bias, whereas too small values of k yield a large variance. For a more detailed discussion on the choice of k we recommend Einmahl et al. (2009).

The proof is based on decomposition (9.21). The first term $\sup_\alpha |\mu_n(R_\alpha^\epsilon) - \mu(R_\alpha^\epsilon)|$ on the right hand side of (9.21) is bounded using Corollary 9.9, while Proposition 9.11 allows to bound the second one (bias term stemming from the tolerance parameter ϵ). Introduce the notation

$$\text{bias}(\alpha, n, k, \epsilon) = 4 \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq \frac{2}{\epsilon}} \left| \frac{n}{k} \tilde{F}_{\alpha, \bar{\alpha}}\left(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z}\right) - g_{\alpha, \bar{\alpha}}(\mathbf{x}, \mathbf{z}) \right|. \quad (9.37)$$

With probability at least $1 - \delta$,

$$\begin{aligned} \forall \emptyset \neq \alpha \subset \{1, \dots, d\}, \\ |\mu_n(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)| \leq Cd \sqrt{\frac{1}{\epsilon k} \log \frac{d+3}{\delta}} + \text{bias}(\alpha, n, k, \epsilon) + Md^2\epsilon. \end{aligned}$$

The upper bound stated in Theorem 9.12 follows.

Remark 9.13. (THRESHOLDING THE ESTIMATOR) In practice, we have to deal with non-asymptotic noisy data, so that many $\widehat{\mathcal{M}}(\alpha)$'s have very small values though the corresponding $\mathcal{M}(\alpha)$'s are null. One solution is thus to define a threshold value, for instance a proportion p of the averaged mass over all the faces α with positive mass, *i.e.* $\text{threshold} = p|A|^{-1} \sum_\alpha \widehat{\mathcal{M}}(\alpha)$ with

$A = \{\alpha, \widehat{\mathcal{M}}(\alpha) > 0\}$. Let us define $\widetilde{\mathcal{M}}(\alpha)$ the obtained thresholded $\widehat{\mathcal{M}}(\alpha)$. Then the estimation error satisfies:

$$\begin{aligned} \|\widetilde{\mathcal{M}} - \mathcal{M}\|_\infty &\leq \|\widetilde{\mathcal{M}} - \widehat{\mathcal{M}}\|_\infty + \|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty \\ &\leq p|A|^{-1} \sum_{\alpha} \widehat{\mathcal{M}}(\alpha) + \|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty \\ &\leq p|A|^{-1} \sum_{\alpha} \mathcal{M}(\alpha) + p|A|^{-1} \sum_{\alpha} |\widehat{\mathcal{M}}(\alpha) - \mathcal{M}(\alpha)| \\ &\quad + \|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty \\ &\leq (p+1)\|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty + p|A|^{-1} \mu([0, 1]^c). \end{aligned}$$

It is outside the scope of this chapter to study optimal values for p . However, Remark 9.14 writes the estimation procedure as an optimization problem, thus exhibiting a link between thresholding and L^1 -regularization.

Remark 9.14. (UNDERLYING RISK MINIMIZATION PROBLEMS) Our estimate $\widehat{\mathcal{M}}(\alpha)$ can be interpreted as a solution of an empirical risk minimization problem inducing a conditional empirical risk \widehat{R}_n . When adding a L^1 regularization term to this problem, we recover $\widetilde{\mathcal{M}}(\alpha)$, the thresholded estimate.

First recall that $\widehat{\mathcal{M}}(\alpha)$ is defined for $\alpha \subset \{1, \dots, d\}$, $\alpha \neq \emptyset$ by $\widehat{\mathcal{M}}(\alpha) = 1/k \sum_{i=1}^n \mathbb{1}_{\frac{k}{n} \widehat{\mathbf{V}}_i \in R_\alpha^c}$. As $R_\alpha^c \subset [0, 1]^c$, we may write

$$\widehat{\mathcal{M}}(\alpha) = \left(\frac{n}{k} \mathcal{P}_n \left(\frac{k}{n} \|\widehat{\mathbf{V}}_1\| \geq 1 \right) \right) \left(\frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}_{\frac{k}{n} \widehat{\mathbf{V}}_i \in R_\alpha^c} \mathbb{1}_{\frac{k}{n} \|\widehat{\mathbf{V}}_i\| \geq 1}}{\mathcal{P}_n(\frac{k}{n} \|\widehat{\mathbf{V}}_1\| \geq 1)} \right),$$

where the last term is the empirical expectation of $Z_{n,i}(\alpha) = \mathbb{1}_{\frac{k}{n} \widehat{\mathbf{V}}_i \in R_\alpha^c}$ conditionnaly to the event $\{\|\frac{k}{n} \widehat{\mathbf{V}}_1\| \geq 1\}$, and $\mathcal{P}_n(\frac{k}{n} \|\widehat{\mathbf{V}}_1\| \geq 1) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\frac{k}{n} \|\widehat{\mathbf{V}}_i\| \geq 1}$. According to Lemma 9.5, for each fixed margin j , $\widehat{V}_i^j \geq \frac{n}{k}$ if, and only if $X_i^j \geq X_{(n-k+1)}^j$, which happens for k observations exactly. Thus,

$$\mathcal{P}_n \left(\frac{k}{n} \|\widehat{\mathbf{V}}_1\| \geq 1 \right) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\exists j, \widehat{V}_i^j \geq \frac{n}{k}} \in \left[\frac{k}{n}, \frac{dk}{n} \right].$$

If we define $\tilde{k} = \tilde{k}(n) \in [k, dk]$ such that $\mathcal{P}_n(\frac{k}{n} \|\widehat{\mathbf{V}}_1\| \geq 1) = \frac{\tilde{k}}{n}$, we then have

$$\begin{aligned} \widehat{\mathcal{M}}(\alpha) &= \frac{\tilde{k}}{k} \left(\frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}_{\frac{k}{n} \widehat{\mathbf{V}}_i \in R_\alpha^c} \mathbb{1}_{\frac{k}{n} \|\widehat{\mathbf{V}}_i\| \geq 1}}{\mathcal{P}_n(\frac{k}{n} \|\widehat{\mathbf{V}}_1\| \geq 1)} \right) \\ &= \frac{\tilde{k}}{k} \arg \min_{m_\alpha > 0} \sum_{i=1}^n (Z_{n,i}(\alpha) - m_\alpha)^2 \mathbb{1}_{\frac{k}{n} \|\widehat{\mathbf{V}}_i\| \geq 1}, \end{aligned}$$

Considering now the $(2^d - 1)$ -vector $\widehat{\mathcal{M}}$ and $\|\cdot\|_{2,\alpha}$ the L^2 -norm on \mathbb{R}^{2^d-1} , we immediatly have (since $k(n)$ does not depend on α)

$$\widehat{\mathcal{M}} = \frac{\tilde{k}}{k} \arg \min_{m \in \mathbb{R}^{2^d-1}} \widehat{R}_n(m), \quad (9.38)$$

where $\widehat{R}_n(m) = \sum_{i=1}^n \|Z_{n,i} - m\|_{2,\alpha}^2 \mathbb{1}_{\frac{k}{n} \|\widehat{\mathbf{V}}_i\| \geq 1}$ is the L^2 -empirical risk of m , restricted to extreme observations, namely to observations \mathbf{X}_i satisfying $\|\widehat{\mathbf{V}}_i\| \geq \frac{n}{k}$. Then, up to a constant $\frac{\tilde{k}}{k} = \Theta(1)$, $\widehat{\mathcal{M}}$ is solution of an empirical conditional risk minimization problem. Define the non-asymptotic theoretical risk $R_n(m)$ for $m \in \mathbb{R}^{2^d-1}$ by

$$R_n(m) = \mathbb{E} \left[\|Z_n - m\|_{2,\alpha}^2 \mathbb{1}_{\left\| \frac{k}{n} \mathbf{V}_1 \right\|_\infty \geq 1} \right]$$

with $Z_n := Z_{n,1}$. Then one can show (see 9.7) that Z_n , conditionally to the event $\{\|\frac{k}{n} \mathbf{V}_1\| \geq 1\}$, converges in distribution to a variable Z_∞ which is a multinomial distribution on \mathbb{R}^{2^d-1} with parameters $(n=1, p_\alpha = \frac{\mu(R_\alpha^\epsilon)}{\mu([\mathbf{0}, \mathbf{1}]^c)}, \alpha \in \{1, \dots, n\}, \alpha \neq \emptyset)$. In other words,

$$\mathbb{P}(Z_\infty(\alpha) = 1) = \frac{\mu(R_\alpha^\epsilon)}{\mu([\mathbf{0}, \mathbf{1}]^c)}$$

for all $\alpha \in \{1, \dots, n\}, \alpha \neq \emptyset$, and $\sum_\alpha Z_\infty(\alpha) = 1$. Thus $R_n(m) \rightarrow R_\infty(m) := \mathbb{E}[\|Z_\infty - m\|_{2,\alpha}^2]$, which is the asymptotic risk. Moreover, the optimization problem

$$\min_{m \in \mathbb{R}^{2^d-1}} R_\infty(m)$$

admits $m = (\frac{\mu(R_\alpha^\epsilon)}{\mu([\mathbf{0}, \mathbf{1}]^c)}, \alpha \in \{1, \dots, n\}, \alpha \neq \emptyset)$ as solution.

Considering the solution of the minimization problem (9.38), which happens to coincide with the definition of $\widehat{\mathcal{M}}$, makes then sense if the goal is to estimate $\mathcal{M} := (\mu(R_\alpha^\epsilon), \alpha \in \{1, \dots, n\}, \alpha \neq \emptyset)$. As well as considering thresholded estimators $\widetilde{\mathcal{M}}(\alpha)$, since it amounts (up to a bias term) to add a L^1 -penalization term to the underlying optimization problem: Let us consider

$$\min_{m \in \mathbb{R}^{2^d-1}} \widehat{R}_n(m) + \lambda \|m\|_{1,\alpha}$$

with $\|m\|_{1,\alpha} = \sum_\alpha |m(\alpha)|$ the L^1 norm on \mathbb{R}^{2^d-1} . In this optimization problem, only extreme observations are involved. It is a well known fact that solving it is equivalent to soft-thresholding the solution of the same problem without the penalty term – and then, up to a bias term due to the **soft**-thresholding, it boils down to setting to zero features $m(\alpha)$ which are less than some fixed threshold $T(\lambda)$. This is an other interpretation on thresholding as defined in Remark 9.13.

9.4 Application to Anomaly Detection

9.4.1 Background on anomaly detection

What is Anomaly Detection ? From a machine learning perspective, anomaly detection can be considered as a specific classification task, where the usual assumption in supervised learning stipulating that the dataset contains structural information regarding all classes breaks down, see Roberts (1999). This typically happens in the case of two highly unbalanced classes: the normal class is expected to regroup a large majority of the dataset, so that the very small number of points representing the abnormal class does not allow to learn information about this class. *Supervised* anomaly detection consists in training the algorithm on a labeled (normal/abnormal) dataset including both normal and abnormal observations. In the *semi-supervised* context, only normal data are available for training. This is the case in applications where normal operations are known but intrusion/attacks/viruses are unknown and should be detected. In the *unsupervised* setup, no assumption is made on the data which consist in unlabeled normal and abnormal instances. In general, a method from the semi-supervised framework may apply to the unsupervised one, as soon as the number of anomalies is sufficiently weak to prevent the algorithm from fitting them when learning the normal behavior. Such a method should be robust to outlying observations.

Extremes and Anomaly Detection. As a matter of fact, ‘extreme’ observations are often more susceptible to be anomalies than others. In other words, extremal observations are often at the *border* between normal and abnormal regions and play a very special role in this context. As the number of observations considered as extreme (*e.g.* in a Peak-over-threshold analysis) typically constitute less than one percent of the data, a classical anomaly detection algorithm would tend to systematically classify all of them as abnormal: it is not worth the risk (in terms of ROC or precision-recall curve for instance) trying to be more accurate in low probability regions without adapted tools. Also, new observations outside the ‘observed support’ are most often predicted as abnormal. However, false positives (*i.e.* false alarms) are very expensive in many applications (*e.g.* aircraft predictive maintenance). It is thus of primal interest to develop tools increasing precision (*i.e.* the probability of observing an anomaly among alarms) on such extremal regions.

Contributions. The algorithm proposed in this chapter provides a scoring function which ranks extreme observations according to their supposed degree of abnormality. This method is complementary to other anomaly detection algorithms, insofar as two algorithms (that described here, together with any other appropriate anomaly detection algorithm) may be trained on the same dataset. Afterwards, the input space may be divided into two regions – an extreme region and a non-extreme one– so that a new observation in the central region (*resp.* in the extremal region) would be classified as abnormal or not according to the scoring function issued by the generic algorithm (*resp.* the one presented here). The scope of our algorithm concerns both semi-supervised and unsupervised problems. Undoubtedly, as it consists in learning a ‘normal’

(i.e. not abnormal) behavior in extremal regions, it is optimally efficient when trained on ‘normal’ observations only. However it also applies to unsupervised situations. Indeed, it involves a non-parametric but relatively coarse estimation scheme which prevents from over-fitting normal data or fitting anomalies. As a consequence, this method is robust to outliers and also applies when the training dataset contains a (small) proportion of anomalies.

9.4.2 DAMEX Algorithm: Detecting Anomalies among Multivariate Extremes

The purpose of this subsection is to explain the heuristic behind the use of multivariate EVT for anomaly detection, which is in fact a natural way to proceed when trying to describe the dependence structure of extreme regions. The algorithm is thus introduced in an intuitive setup, which matches the theoretical framework and results obtained in sections 9.2 and 9.3. The notations are the same as above: $\mathbf{X} = (X^1, \dots, X^d)$ is a random vector in \mathbb{R}^d , with joint (resp. marginal) distribution \mathbf{F} (resp. $F_j, j = 1, \dots, d$) and $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathbf{F}$ is an *i.i.d.* sample. The first natural step to study the dependence between the margins X^j is to standardize them, and the choice of standard Pareto margins (with *c.d.f.* $x \mapsto 1/x$) is convenient: Consider thus the \mathbf{V}_i ’s and $\widehat{\mathbf{V}}_i$ ’s as defined in Section 9.2. One possible strategy to investigate the dependence structure of extreme events is to characterize, for each subset of features $\alpha \subset \{1, \dots, d\}$, the ‘correlation’ of these features given that one of them at least is large and the others are small. Formally, we associate to each such α a coefficient $\mathcal{M}(\alpha)$ reflecting the degree of dependence between the features α . This coefficient is to be proportional to the expected number of points \mathbf{V}_i above a large radial threshold ($\|\mathbf{V}\|_\infty > r$), verifying V_i^j ‘large’ for $j \in \alpha$, while V_i^j ‘small’ for $j \notin \alpha$. In order to define the notion of ‘large’ and ‘small’, fix a (small) tolerance parameter $0 < \epsilon < 1$. Thus, our focus is on the expected proportion of points ‘above a large radial threshold’ r which belong to the truncated rectangles R_α^ϵ defined in (9.10). More precisely, our goal is to estimate the above expected proportion, when the tolerance parameter ϵ goes to 0.

The standard empirical approach –counting the number of points in the regions of interest– leads to estimates $\widehat{\mathcal{M}}(\alpha) = \mu_n(R_\alpha^\epsilon)$ (see (9.19)), with μ_n the empirical version of μ defined in (9.18), namely:

$$\widehat{\mathcal{M}}(\alpha) = \mu_n(R_\alpha^\epsilon) = \frac{n}{k} \widehat{\mathbb{P}}_n \left(\frac{n}{k} R_\alpha^\epsilon \right), \quad (9.39)$$

where we recall that $\widehat{\mathbb{P}}_n = (1/n) \sum_{i=1}^n \delta_{\widehat{\mathbf{V}}_i}$ is the empirical probability distribution of the rank-transformed data, and $k = k(n) > 0$ is such that $k \rightarrow \infty$ and $k = o(n)$ as $n \rightarrow \infty$. The ratio n/k plays the role of a large radial threshold r . From our standardization choice, counting points in $(n/k) R_\alpha^\epsilon$ boils down to selecting, for each feature $j \leq d$, the ‘ k largest values’ X_i^j among n observations. According to the nature of the extremal dependence, a number between k and dk of observations are selected: k in case of perfect dependence, dk in case of ‘independence’, which means, in the EVT framework, that the components may only be large one at a time. In any case, the number of observations considered as extreme is proportional to k , whence the normalizing factor $\frac{n}{k}$.

The coefficients $(\widehat{\mathcal{M}}(\alpha))_{\alpha \in \{1, \dots, d\}}$ associated with the cones \mathcal{C}_α constitute our representation of the dependence structure. This representation is sparse as soon as the $\widehat{\mathcal{M}}(\alpha)$ are positive only for a few groups of features α (compared with the total number of groups, or sub-cones, $2^d - 1$). It is low-dimensional as soon as each of these groups has moderate cardinality $|\alpha|$, *i.e.* as soon as the sub-cones with positive $\widehat{\mathcal{M}}(\alpha)$ are low-dimensional relatively to d .

In fact, up to a normalizing constant, $\widehat{\mathcal{M}}(\alpha)$ is an empirical version of the probability that $T(\mathbf{X})$ belongs to the cone \mathcal{C}_α , conditioned upon exceeding a large threshold. Indeed, for r, n and k sufficiently large, we have (Remark 9.2 and (9.20), reminding that $\mathbf{V} = T(\mathbf{X})$)

$$\widehat{\mathcal{M}}(\alpha) \simeq C \mathbb{P}(T(\mathbf{X}) \in rR_\alpha^\epsilon \mid \|T(\mathbf{X})\| \geq r).$$

Introduce an ‘angular scoring function’

$$w_n(\mathbf{x}) = \sum_{\alpha} \widehat{\mathcal{M}}(\alpha) \mathbb{1}_{\{\widehat{T}(\mathbf{x}) \in R_\alpha^\epsilon\}}. \quad (9.40)$$

For each fixed (new observation) \mathbf{x} , $w_n(\mathbf{x})$ approaches the probability that the random variable \mathbf{X} belongs to the same cone as \mathbf{x} in the transformed space. In short, $w_n(\mathbf{x})$ is an empirical version of the probability that \mathbf{X} and \mathbf{x} have approximately the same ‘direction’. For anomaly detection, the degree of ‘abnormality’ of the new observation \mathbf{x} should be related both to $w_n(\mathbf{x})$ and to the uniform norm $\|\widehat{T}(\mathbf{x})\|_\infty$ (angular and radial components). More precisely, for \mathbf{x} fixed such that $T(\mathbf{x}) \in R_\alpha^\epsilon$. Consider the ‘*directional tail region*’ induced by \mathbf{x} , $A_{\mathbf{x}} = \{\mathbf{y} : T(\mathbf{y}) \in R_\alpha^\epsilon, \|T(\mathbf{y})\|_\infty \geq \|T(\mathbf{x})\|_\infty\}$. Then, if $\|T(\mathbf{x})\|_\infty$ is large enough, we have (using (9.5)) that

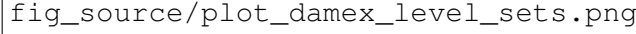
$$\begin{aligned} \mathbb{P}(\mathbf{X} \in A_{\mathbf{x}}) &= \mathbb{P}(\mathbf{V} \in \|T(\mathbf{x})\|_\infty R_\alpha^\epsilon) \\ &= \mathbb{P}(\|\mathbf{V}\| \geq \|T(\mathbf{x})\|) \mathbb{P}(\mathbf{V} \in \|T(\mathbf{x})\|_\infty R_\alpha^\epsilon \mid \|\mathbf{V}\| \geq \|T(\mathbf{x})\|) \\ &\simeq C \mathbb{P}(\|\mathbf{V}\| \geq \|T(\mathbf{x})\|) \widehat{\mathcal{M}}(\alpha) \\ &= C \|\widehat{T}(\mathbf{x})\|_\infty^{-1} w_n(\mathbf{x}). \end{aligned}$$

This yields the scoring function

$$s_n(\mathbf{x}) := \frac{w_n(\mathbf{x})}{\|\widehat{T}(\mathbf{x})\|_\infty}, \quad (9.41)$$

which is thus (up to a scaling constant C) an empirical version of $\mathbb{P}(\mathbf{X} \in A_{\mathbf{x}})$: the smaller $s_n(\mathbf{x})$, the more abnormal the point \mathbf{x} should be considered. As an illustrative example, Figure 9.4 displays the level sets of this scoring function, both in the transformed and the non-transformed input space, in the 2D situation. The data are simulated under a 2D logistic distribution with asymmetric parameters.

This heuristic argument explains the following algorithm, referred to as *Detecting Anomaly with Multivariate EXtremes* (DAMEX in abbreviated form). Note that this is a slightly modified version of the original DAMEX algorithm empirically tested in Goix et al. (2016b), where


FIGURE 9.4: Level sets of s_n on simulated 2D data

ϵ -thickened sub-cones instead of ϵ -thickened rectangles are considered. The proof is more straightforward when considering rectangles and performance remains as good. The complexity is in $O(dn \log n + dn) = O(dn \log n)$, where the first term on the left-hand-side comes from computing the $\hat{F}_j(X_i^j)$ (Step 1) by sorting the data (e.g. merge sort). The second one arises from Step 2.

Algorithm 3 DAMEX

Input: parameters $\epsilon > 0$, $k = k(n)$, $p \geq 0$.

1. Standardize *via* marginal rank-transformation: $\hat{\mathbf{V}}_i := (1/(1 - \hat{F}_j(X_i^j)))_{j=1,\dots,d}$.
2. Assign to each $\hat{\mathbf{V}}_i$ the cone R_α^ϵ it belongs to.
3. Compute $\hat{\mathcal{M}}(\alpha)$ from (9.39) \rightarrow yields: (small number of) cones with non-zero mass.
4. (Optional) Set to 0 the $\hat{\mathcal{M}}(\alpha)$ below some small threshold defined in remark 9.13 *w.r.t.* p . \rightarrow yields: (sparse) representation of the dependence structure

$$\left\{ \hat{\mathcal{M}}(\alpha) : \emptyset \subset \alpha \subset \{1, \dots, d\} \right\}. \quad (9.42)$$

Output: Compute the scoring function given by (9.41),

$$s_n(\mathbf{x}) := (1/\|\hat{T}(\mathbf{x})\|_\infty) \sum_{\alpha} \hat{\mathcal{M}}(\alpha) \mathbb{1}_{\hat{T}(\mathbf{x}) \in R_\alpha^\epsilon}.$$

Before investigating how the algorithm above empirically performs when applied to synthetic/real datasets, a few remarks are in order.

Remark 9.15. (INTERPRETATION OF THE PARAMETERS) In view of (9.39), n/k is the threshold above which the data are considered as extreme and k is proportional to the number of such data, a common approach in multivariate extremes. The tolerance parameter ϵ accounts for the non-asymptotic nature of data. The smaller k , the smaller ϵ shall be chosen. The additional angular mass threshold in step 4. acts as an additional sparsity inducing parameter. Note that even

without this additional step (*i.e.* setting $p = 0$, the obtained representation for real-world data (see Table 9.2) is already sparse (the number of charges cones is significantly less than 2^d).

Remark 9.16. (CHOICE OF PARAMETERS) A standard choice of parameters (ϵ, k, p) is respectively $(0.01, n^{1/2}, 0.1)$. However, there is no simple manner to choose optimally these parameters, as there is no simple way to determine how fast is the convergence to the (asymptotic) extreme behavior –namely how far in the tail appears the asymptotic dependence structure. Indeed, even though the first term of the error bound in Theorem 9.12 is proportional, up to re-scaling, to $\sqrt{\frac{1}{\epsilon k}} + \sqrt{\epsilon}$, which suggests choosing ϵ of order $k^{-1/4}$, the unknown bias term perturbs the analysis and in practice, one obtains better results with the values above mentioned. In a supervised or semi-supervised framework (or if a small labeled dataset is available) these three parameters should be chosen by cross-validation. In the unsupervised situation, a classical heuristic (Coles (2001)) is to choose (k, ϵ) in a stability region of the algorithm’s output: the largest k (*resp.* the larger ϵ) such that when decreased, the dependence structure remains stable. This amounts to selecting as many data as possible as being extreme (*resp.* in low dimensional regions), within a stability domain of the estimates, which exists under the primal assumption (9.1) and in view of Lemma 9.1.

Remark 9.17. (DIMENSION REDUCTION) If the extreme dependence structure is low dimensional, namely concentrated on low dimensional cones \mathcal{C}_α – or in other terms if only a limited number of margins can be large together – then most of the \hat{V}_i ’s will be concentrated on the R_α^ϵ ’s such that $|\alpha|$ (the dimension of the cone \mathcal{C}_α) is small; then the representation of the dependence structure in (9.42) is both sparse and low dimensional.

Remark 9.18. (SCALING INVARIANCE) DAMEX produces the same result if the input data are transformed in such a way that the marginal order is preserved. In particular, any marginally increasing transform or any scaling as a preprocessing step does not affect the algorithm. It also implies invariance with respect to any change in the measuring units. This invariance property constitutes part of the strength of the algorithm, since data preprocessing steps usually have a great impact on the overall performance and are of major concern in practice.

9.5 Experimental results

9.5.1 Recovering the support of the dependence structure of generated data

Datasets of size 50000 (respectively 100000, 150000) are generated in \mathbb{R}^{10} according to a popular multivariate extreme value model, introduced by Tawn (1990), namely a multivariate asymmetric logistic distribution (G_{log}). The data have the following features: (i) they resemble ‘real life’ data, that is, the X_i^j ’s are non zero and the transformed \hat{V}_i ’s belong to the interior cone $\mathcal{C}_{\{1, \dots, d\}}$, (ii) the associated (asymptotic) exponent measure concentrates on K disjoint cones $\{\mathcal{C}_{\alpha_m}, 1 \leq m \leq K\}$. For the sake of reproducibility, $G_{log}(\mathbf{x}) = \exp\{-\sum_{m=1}^K \left(\sum_{j \in \alpha_m} (|A(j)|x_j)^{-1/w_{\alpha_m}}\right)^{w_{\alpha_m}}\}$, where $|A(j)|$ is the cardinal of the set $\{\alpha \in D : j \in \alpha\}$ and where $w_{\alpha_m} = 0.1$ is a dependence

parameter (strong dependence). The data are simulated using Algorithm 2.2 in Stephenson (2003). The subset of sub-cones D charged by μ is randomly chosen (for each fixed number of sub-cones K) and the purpose is to recover D by Algorithm 3. For each K , 100 experiments are made and we consider the number of ‘errors’, that is, the number of non-recovered or false-discovered sub-cones. Table 9.1 shows the averaged numbers of errors among the 100 experiments. The results are very promising in situations where the number of sub-cones is

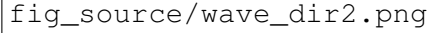
# sub-cones K	3	5	10	15	20	25	30	35	40	45	50
Aver. # errors (n=5e4)	0.02	0.65	0.95	0.45	0.49	1.35	4.19	8.9	15.46	19.92	18.99
Aver. # errors (n=10e4)	0.00	0.45	0.36	0.21	0.13	0.43	0.38	0.55	1.91	1.67	2.37
Aver. # errors (n=15e4)	0.00	0.34	0.47	0.00	0.02	0.13	0.13	0.31	0.39	0.59	1.77

TABLE 9.1: Support recovering on simulated data

moderate *w.r.t.* the number of observations.

9.5.2 Sparse structure of extremes (wave data)

Our goal is here to verify that the two expected phenomena mentioned in the introduction, **1-** sparse dependence structure of extremes (small number of sub-cones with non zero mass), **2-** low dimension of the sub-cones with non-zero mass, do occur with real data. We consider wave directions data provided by Shell, which consist of 58585 measurements D_i , $i \leq 58595$ of wave directions between 0° and 360° at 50 different locations (buoys in North sea). The dimension is thus 50. The angle 90° being fairly rare, we work with data obtained as $X_i^j = 1/(10^{-10} + |90 - D_i^j|)$, where D_i^j is the wave direction at buoy j , time i . Thus, D_i^j 's close to 90 correspond to extreme X_i^j 's. Results in Table 9.2 show that the number of sub-cones \mathcal{C}_α identified by Algorithm 3 is indeed small compared to the total number of sub-cones ($2^{50}-1$). (Phenomenon **1** in the introduction section). Further, the dimension of these sub-cones is essentially moderate (Phenomenon **2**): respectively 93%, 98.6% and 99.6% of the mass is affected to sub-cones of dimension no greater than 10, 15 and 20 respectively (to be compared with $d = 50$). Histograms displaying the mass repartition produced by Algorithm 3 are given in Fig. 9.5.



fig_source/wave_dir2.png

FIGURE 9.5: sub-cone dimensions of wave data

	non-extreme data	extreme data
nb of sub-cones with mass > 0 ($p = 0$)	3413	858
idem after thresholding ($p = 0.1$)	2	64
idem after thresholding ($p = 0.2$)	1	18

TABLE 9.2: Total number of sub-cones of wave data

9.5.3 Application to Anomaly Detection on real-world data sets

The main purpose of Algorithm 3 is to build a ‘normal profile’ for extreme data, so as to distinguish between normal and ab-normal extremes. In this section we evaluate its performance and compare it with that of a standard anomaly detection algorithm, the Isolation Forest (iForest) algorithm, which we chose in view of its established high performance (Liu et al. (2008)). The two algorithms are trained and tested on the same datasets, the test set being restricted to an extreme region. Five reference anomaly detection datasets are considered: *shuttle*, *forestcover*, *http*, *SF* and *SA* ¹. The experiments are performed in a semi-supervised framework (the training set consists of normal data).

The *shuttle* dataset is the fusion of the training and testing datasets available in the UCI repository Lichman (2013). The data have 9 numerical attributes, the first one being time. Labels from 7 different classes are also available. Class 1 instances are considered as normal, the others as anomalies. We use instances from all different classes but class 4, which yields an anomaly ratio (class 1) of 7.17%.

In the *forestcover* data, also available at UCI repository (Lichman (2013)), the normal data are the instances from class 2 while instances from class 4 are anomalies, other classes are omitted, so that the anomaly ratio for this dataset is 0.9%.

¹These datasets are available for instance on <http://scikit-learn.org/dev/>

The last three datasets belong to the KDD Cup '99 dataset (KDDCup (1999), Tavallae et al. (2009)), produced by processing the tcpdump portions of the 1998 DARPA Intrusion Detection System (IDS) Evaluation dataset, created by MIT Lincoln Lab Lippmann et al. (2000). The artificial data was generated using a closed network and a wide variety of hand-injected attacks (anomalies) to produce a large number of different types of attack with normal activity in the background. Since the original demonstrative purpose of the dataset concerns supervised anomaly detection, the anomaly rate is very high (80%), which is unrealistic in practice, and inappropriate for evaluating the performance on realistic data. We thus take standard pre-processing steps in order to work with smaller anomaly rates. For datasets *SF* and *http* we proceed as described in Yamanishi et al. (2000): *SF* is obtained by picking up the data with positive logged-in attribute, and focusing on the intrusion attack, which gives an anomaly proportion of 0.48%. The dataset *http* is a subset of *SF* corresponding to a third feature equal to 'http'. Finally, the *SA* dataset is obtained as in Eskin et al. (2002) by selecting all the normal data, together with a small proportion (1%) of anomalies.

Table 9.3 summarizes the characteristics of these datasets. The thresholding parameter p is fixed to 0.1, the averaged mass of the non-empty sub-cones, while the parameters (k, ϵ) are standardly chosen as $(n^{1/2}, 0.01)$. The extreme region on which the evaluation step is performed is chosen as $\{\mathbf{x} : \|T(\mathbf{x})\| > \sqrt{n}\}$, where n is the training set's sample size. The ROC and PR curves are computed using only observations in the extreme region. This provides a precise evaluation of the two anomaly detection methods on extreme data. For each of them, 20 experiments on random training and testing datasets are performed, yielding averaged ROC and Precision-Recall curves whose AUC are presented in Table 9.4. DAMEX significantly improves the performance (both in term of precision and of ROC curves) in extreme regions for each dataset, as illustrated in figures 9.6 and 9.7.

In Table 9.5, we repeat the same experiments but with $\epsilon = 0.1$. This yields the same strong performance of DAMEX, excepting for *SF* (see Figure 9.8). Generally, too large ϵ may yield over-estimated $\widehat{\mathcal{M}}(\alpha)$ for low-dimensional faces α . Such a performance gap between $\epsilon = 0.01$ and $\epsilon = 0.1$ can also be explained by the fact that anomalies may form a cluster which is wrongly include in some over-estimated 'normal' sub-cone, when ϵ is too large. Such singular anomaly structure would also explain the counter performance of iForest on this dataset.

We also point out that for very small values of epsilon ($\epsilon \leq 0.001$), the performance of DAMEX significantly decreases on these datasets. With such a small ϵ , most observations belong to the central cone (the one of dimension d) which is widely over-estimated, while the other cones are under-estimated.

The only case where using very small ϵ should be useful, is when the asymptotic behaviour is clearly reached at level k (usually for very large threshold n/k , e.g. $k = n^{1/3}$), or in the specific case where anomalies clearly concentrate in low dimensional sub-cones: The use of a small ϵ precisely allows to assign a high abnormality score to these subcones (under-estimation of the asymptotic mass), which yields better performances.

The averaged ROC curves and PR curves for the other datasets are represented in Figures

	shuttle	forestcover	SA	SF	http
Samples total	85849	286048	976158	699691	619052
Number of features	9	54	41	4	3
Percentage of anomalies	7.17	0.96	0.35	0.48	0.39

TABLE 9.3: Datasets characteristics

Dataset	iForest		DAMEX	
	AUC ROC	AUC PR	AUC ROC	AUC PR
shuttle	0.957	0.987	0.988	0.996
forestcover	0.667	0.201	0.976	0.805
http	0.561	0.321	0.981	0.742
SF	0.134	0.189	0.988	0.973
SA	0.932	0.625	0.945	0.818

TABLE 9.4: Results on extreme regions with standard parameters $(k, \epsilon) = (n^{1/2}, 0.01)$

Considering the significant performance improvements on extreme data, DAMEX may be combined with any standard anomaly detection algorithm to handle extreme *and* non-extreme data. This would improve the *global* performance of the chosen standard algorithm, and in particular decrease the false alarm rate (increase the slope of the ROC curve's tangents near the origin). This combination can be done by splitting the input space between an extreme region and a non-extreme one, then using Algorithm 3 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.

9.6 Conclusion

The contribution of this chapter is twofold. First, it brings advances in multivariate EVT by designing a statistical method that possibly exhibits a sparsity pattern in the dependence structure of extremes, while deriving non-asymptotic bounds to assess the accuracy of the estimation


Dataset	iForest		DAMEX	
	AUC ROC	AUC PR	AUC ROC	AUC PR
shuttle	0.957	0.987	0.980	0.995
forestcover	0.667	0.201	0.984	0.852
http	0.561	0.321	0.971	0.639
SF	0.134	0.189	0.101	0.211
SA	0.932	0.625	0.964	0.848

TABLE 9.5: Results on extreme regions with lower $\epsilon = 0.1$


fig_source/shuttle-semi-supervised-average-rect-01.png

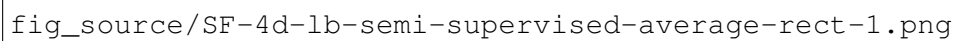
FIGURE 9.6: SF dataset, default parameters

procedure. Our method is intended to be used as a preprocessing step to scale up multivariate extreme values modeling to high dimensional settings, which is currently one of the major challenges in multivariate EVT. Since the asymptotic bias ($\text{bias}(\alpha, n, k, \epsilon)$ in eq. (9.37)) appears as a separate term in the bound established, no second order assumption is required. One possible line of further research would be to make such an assumption (*i.e.* to assume that the bias itself is regularly varying), in order to choose ϵ adaptively with respect to k and n (see Remark 9.16). This might also open up the possibility of de-biasing the estimation procedure (Fougeres et al. (2015), Beirlant et al. (2015)). As a second contribution, this work extends the applicability of



fig_source/SF-4d-lb-semi-supervised-average-rect-01.png

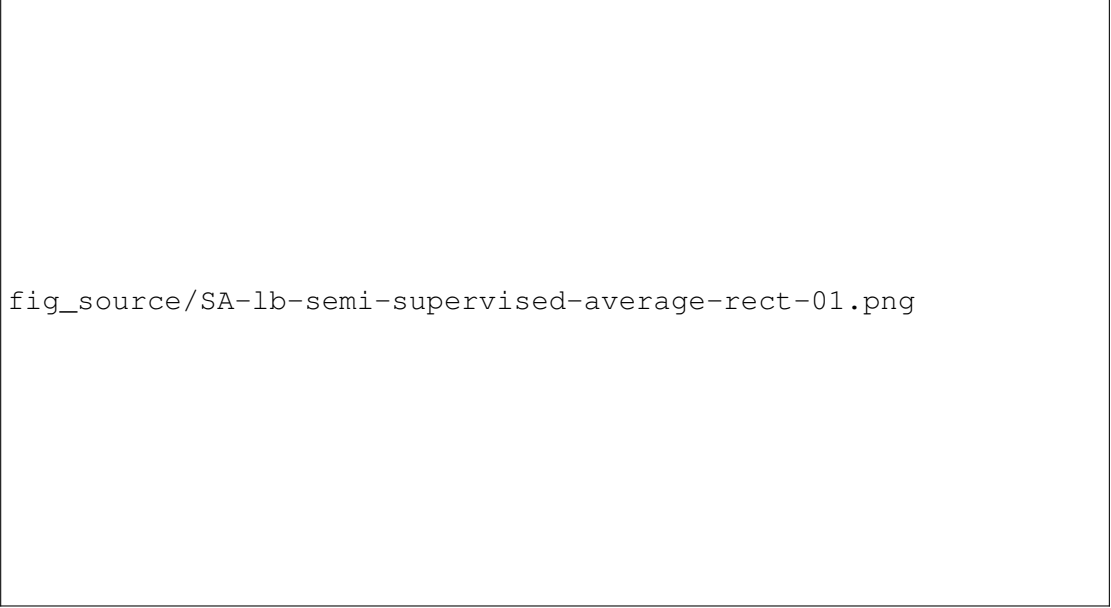
FIGURE 9.7: SF dataset, larger ϵ



fig_source/SF-4d-lb-semi-supervised-average-rect-1.png


FIGURE 9.8: SF dataset, larger ϵ

multivariate EVT to the field of anomaly detection: a multivariate EVT-based algorithm which scores extreme observations according to their degree of abnormality is proposed. Due to its moderate complexity –of order $dn \log n$ – this algorithm is suitable for the treatment of real word large-scale learning problems, and experimental results reveal a significantly increased performance on extreme regions compared with standard anomaly detection approaches.



fig_source/SA-lb-semi-supervised-average-rect-01.png

FIGURE 9.9: SA dataset, default parameters



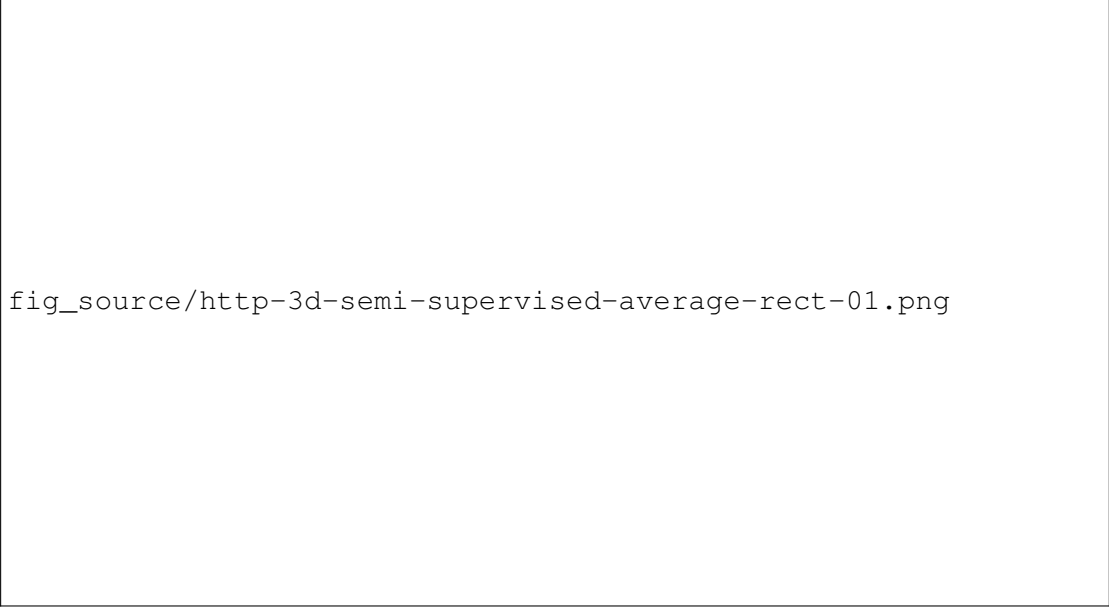
fig_source/forestcover-semi-supervised-average-rect-01.png

FIGURE 9.10: forestcover dataset, default parameters

9.7 Technical proofs

9.7.1 Proof of Lemma 9.5

For n vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ in \mathbb{R}^d , let us denote by $\text{rank}(v_i^j)$ the rank of v_i^j among v_1^j, \dots, v_n^j , that is $\text{rank}(v_i^j) = \sum_{k=1}^n \mathbb{1}_{\{v_k^j \leq v_i^j\}}$, so that $\hat{F}_j(X_i^j) = (\text{rank}(X_i^j) - 1)/n$. For the first equivalence,



fig_source/http-3d-semi-supervised-average-rect-01.png

FIGURE 9.11: http dataset, default parameters

notice that $\hat{V}_i^j = 1/\hat{U}_i^j$. For the others, we have both at the same time:

$$\begin{aligned}
 \hat{V}_i^j \geq \frac{n}{k} x_j &\Leftrightarrow 1 - \frac{\text{rank}(X_i^j) - 1}{n} \leq \frac{k}{n} x_j^{-1} \\
 &\Leftrightarrow \text{rank}(X_i^j) \geq n - kx_j^{-1} + 1 \\
 &\Leftrightarrow \text{rank}(X_i^j) \geq n - \lfloor kx_j^{-1} \rfloor + 1 \\
 &\Leftrightarrow X_i^j \geq X_{(n - \lfloor kx_j^{-1} \rfloor + 1)}^j,
 \end{aligned}$$

and

$$\begin{aligned}
 X_i^j \geq X_{(n - \lfloor kx_j^{-1} \rfloor + 1)}^j &\Leftrightarrow \text{rank}(X_i^j) \geq n - \lfloor kx_j^{-1} \rfloor + 1 \\
 &\Leftrightarrow \text{rank}(F_j(X_i^j)) \geq n - \lfloor kx_j^{-1} \rfloor + 1 \quad (\text{with probability one}) \\
 &\Leftrightarrow \text{rank}(1 - F_j(X_i^j)) \leq \lfloor kx_j^{-1} \rfloor \\
 &\Leftrightarrow U_i^j \leq U_{(\lfloor kx_j^{-1} \rfloor)}^j.
 \end{aligned}$$

9.7.2 Proof of Lemma 9.6

First, recall that $g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) = \mu(R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta))$, see (9.29). Denote by π the transformation to pseudo-polar coordinates introduced in Section 9.2,

$$\begin{aligned}
 \pi : [0, \infty]^d \setminus \{\mathbf{0}\} &\rightarrow (0, \infty] \times S_{\infty}^{d-1} \\
 \mathbf{v} &\mapsto (r, \boldsymbol{\theta}) = (\|\mathbf{v}\|_{\infty}, \|\mathbf{v}\|_{\infty}^{-1} \mathbf{v}).
 \end{aligned}$$

Then, we have $d(\mu \circ \pi^{-1}) = \frac{dr}{r^2} d\Phi$ on $(0, \infty] \times S_{\infty}^{d-1}$. This classical result from EVT comes from the fact that, for $r_0 > 0$ and $B \subset S_{\infty}^{d-1}$, $\mu \circ \pi^{-1}\{r \geq r_0, \boldsymbol{\theta} \in B\} = r_0^{-1} \Phi(B)$, see (9.5). Then

$$\begin{aligned}
g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) &= \mu \circ \pi^{-1} \left\{ (r, \boldsymbol{\theta}) : \quad \forall i \in \alpha, r\theta_i \geq x_i^{-1}; \quad \forall j \in \beta, r\theta_j < z_j^{-1} \right\} \\
&= \mu \circ \pi^{-1} \left\{ (r, \boldsymbol{\theta}) : \quad r \geq \bigvee_{i \in \alpha} (\theta_i x_i)^{-1}; \quad r < \bigwedge_{j \in \beta} (\theta_j z_j)^{-1} \right\} \\
&= \int_{\boldsymbol{\theta} \in S_{\infty}^{d-1}} \int_{r>0} \mathbb{1}_{r \geq \bigvee_{i \in \alpha} (\theta_i x_i)^{-1}} \mathbb{1}_{r < \bigwedge_{j \in \beta} (\theta_j z_j)^{-1}} \frac{dr}{r^2} d\Phi(\boldsymbol{\theta}) \\
&= \int_{\boldsymbol{\theta} \in S_{\infty}^{d-1}} \left(\left(\bigvee_{i \in \alpha} (\theta_i x_i)^{-1} \right)^{-1} - \left(\bigwedge_{j \in \beta} (\theta_j z_j)^{-1} \right)^{-1} \right) d\Phi(\boldsymbol{\theta}) \\
&= \int_{\boldsymbol{\theta} \in S_{\infty}^{d-1}} \left(\bigwedge_{i \in \alpha} \theta_i x_i - \bigvee_{j \in \beta} \theta_j z_j \right) d\Phi(\boldsymbol{\theta}),
\end{aligned}$$

which proves the first assertion. To prove the Lipschitz property, notice first that, for any finite sequence of real numbers c and d , $\max_i c_i - \max_i d_i \leq \max_i (c_i - d_i)$ and $\min_i c_i - \min_i d_i \leq \max_i (c_i - d_i)$. Thus for every $\mathbf{x}, \mathbf{z} \in [0, \infty]^d \setminus \{\infty\}$ and $\boldsymbol{\theta} \in S_{\infty}^{d-1}$:

$$\begin{aligned}
&\left(\bigwedge_{j \in \alpha} \theta_j x_j - \bigvee_{j \in \beta} \theta_j z_j \right) - \left(\bigwedge_{j \in \alpha} \theta_j x'_j - \bigvee_{j \in \beta} \theta_j z'_j \right) \\
&\leq \left[\left(\bigwedge_{j \in \alpha} \theta_j x_j - \bigvee_{j \in \beta} \theta_j z_j \right) - \left(\bigwedge_{j \in \alpha} \theta_j x'_j - \bigvee_{j \in \beta} \theta_j z'_j \right) \right] \\
&\leq \left[\bigwedge_{j \in \alpha} \theta_j x_j - \bigwedge_{j \in \alpha} \theta_j x'_j + \bigvee_{j \in \beta} \theta_j z'_j - \bigvee_{j \in \beta} \theta_j z_j \right] \\
&\leq \left[\max_{j \in \alpha} (\theta_j x_j - \theta_j x'_j) + \max_{j \in \beta} (\theta_j z'_j - \theta_j z_j) \right] \\
&\leq \max_{j \in \alpha} \theta_j |x_j - x'_j| + \max_{j \in \beta} \theta_j |z'_j - z_j|
\end{aligned}$$

Hence,

$$\begin{aligned}
&|g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) - g_{\alpha, \beta}(\mathbf{x}', \mathbf{z}')| \\
&\leq \int_{S_{\infty}^{d-1}} \left(\max_{j \in \alpha} \theta_j |x_j - x'_j| + \max_{j \in \beta} \theta_j |z'_j - z_j| \right) d\Phi(\boldsymbol{\theta}).
\end{aligned}$$

Now, by (9.6) we have:

$$\int_{S_{\infty}^{d-1}} \max_{j \in \alpha} \theta_j |x_j - x'_j| d\Phi(\boldsymbol{\theta}) = \mu([0, \tilde{\mathbf{x}}^{-1}]^c)$$

with $\tilde{\mathbf{x}}$ defined as $\tilde{x}_j = |x_j - x'_j|$ for $j \in \alpha$, and 0 elsewhere. It suffices then to write:

$$\begin{aligned} \mu([\mathbf{0}, \tilde{\mathbf{x}}^{-1}]^c) &= \mu(\{y, \exists j \in \alpha, y_j \geq |x_j - x'_j|^{-1}\}) \\ &\leq \sum_{j \in \alpha} \mu(\{y, y_j \geq |x_j - x'_j|^{-1}\}) \\ &\leq \sum_{j \in \alpha} |x_j - x'_j|. \end{aligned}$$

Similarly, $\int_{S_\infty^{d-1}} \max_{j \in \beta} \theta_j |z'_j - z_j| \, d\Phi(\boldsymbol{\theta}) \leq \sum_{j \in \beta} |z_j - z'_j|$.

9.7.3 Proof of Proposition 9.8

The starting point is inequality (9) on p.7 in Goix et al. (2015b) which bounds the deviation of the empirical measure on extreme regions. Let $\mathcal{C}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{\mathbf{Z}_i \in \cdot\}}$ and $\mathcal{C}(\mathbf{x}) = \mathbb{P}(\mathbf{Z} \in \cdot)$ be the empirical and true measures associated with a n -sample $\mathbf{Z}_1, \dots, \mathbf{Z}_d$ of *i.i.d.* realizations of a random vector $\mathbf{Z} = (Z^1, \dots, Z^d)$ with uniform margins on $[0, 1]$. Then for any real number $\delta \geq e^{-k}$, with probability greater than $1 - \delta$,

$$\sup_{0 \leq \mathbf{x} \leq T} \frac{n}{k} \left| \mathcal{C}_n\left(\frac{k}{n}[\mathbf{x}, \infty]^c\right) - \mathcal{C}\left(\frac{k}{n}[\mathbf{x}, \infty]^c\right) \right| \leq Cd \sqrt{\frac{T}{k} \log \frac{1}{\delta}}. \quad (9.43)$$

Recall that with the above notations, $0 \leq \mathbf{x} \leq T$ means $0 \leq x_j \leq T$ for every j . The proof of Proposition 9.8 follows the same lines as in Goix et al. (2015b). The cornerstone concentration inequality (9.43) has to be replaced with

$$\begin{aligned} \max_{\alpha, \beta} \sup_{\substack{0 \leq \mathbf{x}, \mathbf{z} \leq T \\ \exists j \in \alpha, x_j \leq T'}} \frac{n}{k} \left| \mathcal{C}_n\left(\frac{k}{n}R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta)^{-1}\right) - \mathcal{C}\left(\frac{k}{n}R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta)^{-1}\right) \right| \\ \leq Cd \sqrt{\frac{dT'}{k} \log \frac{1}{\delta}}. \end{aligned} \quad (9.44)$$

Remark 9.19. Inequality (9.44) is here written in its full generality, namely with a separate constant T' possibly smaller than T . If $T' < T$, we then have a smaller bound (typically, we may use $T = 1/\epsilon$ and $T' = 1$). However, we only use (9.44) with $T = T'$ in the analysis below, since the smaller bounds in T' obtained (on $\Lambda(n)$ in (9.47)) would be diluted (by $\Upsilon(n)$ in (9.47)).

Proof of (9.44). Recall that for notational convenience we write ‘ α, β ’ for ‘ α non-empty subset of $\{1, \dots, d\}$ and β subset of $\{1, \dots, d\}$ ’. The key is to apply Theorem 1 in Goix et al. (2015b),

with a VC-class which fits our purposes. Namely, consider

$$\begin{aligned}\mathcal{A} &= \mathcal{A}_{T,T'} = \bigcup_{\alpha,\beta} \mathcal{A}_{T,T',\alpha,\beta} \quad \text{with} \\ \mathcal{A}_{T,T',\alpha,\beta} &= \frac{k}{n} \left\{ R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta)^{-1} : \mathbf{x}, \mathbf{z} \in \mathbb{R}^d, 0 \leq \mathbf{x}, \mathbf{z} \leq T, \right. \\ &\quad \left. \exists j \in \alpha, x_j \leq T' \right\},\end{aligned}$$

for $T, T' > 0$ and $\alpha, \beta \subset \{1, \dots, d\}$, $\alpha \neq \emptyset$. \mathcal{A} has VC-dimension $V_{\mathcal{A}} = d$, as the one considered in Goix et al. (2015b). Recall in view of (9.26) that

$$\begin{aligned}R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta)^{-1} &= \left\{ \mathbf{y} \in [0, \infty]^d, y_j \leq x_j \text{ for } j \in \alpha, \right. \\ &\quad \left. y_j > z_j \text{ for } j \in \beta \right\} \\ &= [\mathbf{a}, \mathbf{b}],\end{aligned}$$

with \mathbf{a} and \mathbf{b} defined by $a_j = \begin{cases} 0 & \text{for } j \in \alpha \\ z_j & \text{for } j \in \beta \end{cases}$ and $b_j = \begin{cases} x_j & \text{for } j \in \alpha \\ \infty & \text{for } j \in \beta \end{cases}$. Since we have

$\forall A \in \mathcal{A}, A \subset [\frac{k}{n}\mathbf{T}', \infty[^c$, the probability for a r.v. \mathbf{Z} with uniform margins in $[0, 1]$ to be in the union class $\mathbb{A} = \bigcup_{A \in \mathcal{A}} A$ is $\mathbb{P}(\mathbf{Z} \in \mathbb{A}) \leq \mathbb{P}(\mathbf{Z} \in [\frac{k}{n}\mathbf{T}', \infty[^c) \leq \sum_{j=1}^d \mathbb{P}(Z^j \leq \frac{k}{n}T') \leq \frac{k}{n}dT'$. Inequality (9.44) is thus a direct consequence of Theorem 1 in Goix et al. (2015b). \square

Define now the empirical version $\tilde{F}_{n,\alpha,\beta}$ of $\tilde{F}_{\alpha,\beta}$ (introduced in (9.28)) as

$$\tilde{F}_{n,\alpha,\beta}(\mathbf{x}, \mathbf{z}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i^j \leq x_j \text{ for } j \in \alpha \text{ and } U_i^j > z_j \text{ for } j \in \beta\}}, \quad (9.45)$$

so that $\frac{n}{k} \tilde{F}_{n,\alpha,\beta}(\frac{k}{n}\mathbf{x}, \frac{k}{n}\mathbf{z}) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{\{U_i^j \leq \frac{k}{n}x_j \text{ for } j \in \alpha \text{ and } U_i^j > \frac{k}{n}z_j \text{ for } j \in \beta\}}$. Notice that the U_i^j 's are not observable (since F_j is unknown). In fact, $\tilde{F}_{n,\alpha,\beta}$ will be used as a substitute for $g_{n,\alpha,\beta}$ (defined in (9.30)) allowing to handle uniform variables. This is illustrated by the following lemmas.

Lemma 9.20 (Link between $g_{n,\alpha,\beta}$ and $\tilde{F}_{n,\alpha,\beta}$). *The empirical version of $\tilde{F}_{\alpha,\beta}$ and that of $g_{\alpha,\beta}$ are related via*

$$g_{n,\alpha,\beta}(\mathbf{x}, \mathbf{z}) = \frac{n}{k} \tilde{F}_{n,\alpha,\beta} \left(\left(U_{(\lfloor kx_j \rfloor)}^j \right)_{j \in \alpha}, \left(U_{(\lfloor kz_j \rfloor)}^j \right)_{j \in \beta} \right),$$

Proof. Considering the definition in (9.45) and (9.31), both sides are equal to $\mu_n(R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta))$. \blacksquare

\square

Lemma 9.21 (Uniform bound on $\tilde{F}_{n,\alpha,\beta}$'s deviations). *For any finite $T > 0$, and $\delta \geq e^{-k}$, with probability at least $1 - \delta$, the deviation of $\tilde{F}_{n,\alpha,\beta}$ from $\tilde{F}_{\alpha,\beta}$ is uniformly bounded:*

$$\max_{\alpha,\beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} \left| \frac{n}{k} \tilde{F}_{n,\alpha,\beta} \left(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z} \right) - \frac{n}{k} \tilde{F}_{\alpha,\beta} \left(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z} \right) \right| \leq Cd \sqrt{\frac{T}{k} \log \frac{1}{\delta}}.$$

Proof. Notice that

$$\begin{aligned} & \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} \left| \frac{n}{k} \tilde{F}_{n,\alpha,\beta} \left(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z} \right) - \frac{n}{k} \tilde{F}_{\alpha,\beta} \left(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z} \right) \right| \\ &= \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} \frac{n}{k} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{U}_i \in \frac{k}{n} R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta)^{-1}} - \mathbb{P} \left[\mathbf{U} \in \frac{k}{n} R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta)^{-1} \right] \right|, \end{aligned}$$

and apply inequality (9.44) with $T' = T$. \square

Remark 9.22. Note that the following stronger inequality holds true, when using (9.44) in full generality, i.e. with $T' < T$. For any finite $T, T' > 0$, and $\delta \geq e^{-k}$, with probability at least $1 - \delta$,

$$\max_{\alpha,\beta} \sup_{\substack{0 \leq \mathbf{x}, \mathbf{z} \leq T \\ \exists j \in \alpha, x_j \leq T'}} \left| \frac{n}{k} \tilde{F}_{n,\alpha,\beta} \left(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z} \right) - \frac{n}{k} \tilde{F}_{\alpha,\beta} \left(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z} \right) \right| \leq Cd \sqrt{\frac{T'}{k} \log \frac{1}{\delta}}.$$

The following lemma is stated and proved in Goix et al. (2015b).

Lemma 9.23 (Bound on the order statistics of \mathbf{U}). *Let $\delta \geq e^{-k}$. For any finite positive number $T > 0$ such that $T \geq 7/2((\log d)/k + 1)$, we have with probability greater than $1 - \delta$,*

$$\forall 1 \leq j \leq d, \quad \frac{n}{k} U_{(\lfloor kT \rfloor)}^j \leq 2T, \quad (9.46)$$

and with probability greater than $1 - (d + 1)\delta$,

$$\max_{1 \leq j \leq d} \sup_{0 \leq x_j \leq T} \left| \frac{\lfloor kx_j \rfloor}{k} - \frac{n}{k} U_{(\lfloor kx_j \rfloor)}^j \right| \leq C \sqrt{\frac{T}{k} \log \frac{1}{\delta}}.$$

We may now proceed with the proof of Proposition 9.8. Using Lemma 9.20, we may write:

$$\begin{aligned} & \max_{\alpha,\beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} |g_{n,\alpha,\beta}(\mathbf{x}, \mathbf{z}) - g_{\alpha,\beta}(\mathbf{x}, \mathbf{z})| \\ &= \max_{\alpha,\beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} \left| \frac{n}{k} \tilde{F}_{n,\alpha,\beta} \left(\left(U_{(\lfloor kx_j \rfloor)}^j \right)_{j \in \alpha}, \left(U_{(\lfloor kz_j \rfloor)}^j \right)_{j \in \beta} \right) - g_{\alpha,\beta}(\mathbf{x}, \mathbf{z}) \right| \\ &\leq \Lambda(n) + \Xi(n) + \Upsilon(n). \end{aligned} \quad (9.47)$$

with:

$$\begin{aligned}
\Lambda(n) &= \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} \left| \frac{n}{k} \tilde{F}_{n, \alpha, \beta} \left(\left(U_{(\lfloor kx_j \rfloor)}^j \right)_{j \in \alpha}, \left(U_{(\lfloor kz_j \rfloor)}^j \right)_{j \in \beta} \right) \right. \\
&\quad \left. - \frac{n}{k} \tilde{F}_{\alpha, \beta} \left(\left(U_{(\lfloor kx_j \rfloor)}^j \right)_{j \in \alpha}, \left(U_{(\lfloor kz_j \rfloor)}^j \right)_{j \in \beta} \right) \right| \\
\Xi(n) &= \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} \left| \frac{n}{k} \tilde{F}_{\alpha, \beta} \left(\left(U_{(\lfloor kx_j \rfloor)}^j \right)_{j \in \alpha}, \left(U_{(\lfloor kz_j \rfloor)}^j \right)_{j \in \beta} \right) \right. \\
&\quad \left. - g_{\alpha, \beta} \left(\left(\frac{n}{k} U_{(\lfloor kx_j \rfloor)}^j \right)_{j \in \alpha}, \left(\frac{n}{k} U_{(\lfloor kz_j \rfloor)}^j \right)_{j \in \beta} \right) \right| \\
\Upsilon(n) &= \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} \left| g_{\alpha, \beta} \left(\left(\frac{n}{k} U_{(\lfloor kx_j \rfloor)}^j \right)_{j \in \alpha}, \left(\frac{n}{k} U_{(\lfloor kz_j \rfloor)}^j \right)_{j \in \beta} \right) - g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) \right|.
\end{aligned}$$

Now, considering (9.46) we have with probability greater than $1 - \delta$ that for every $1 \leq j \leq d$, $U_{(\lfloor kT \rfloor)}^j \leq 2T \frac{k}{n}$, so that

$$\Lambda(n) \leq \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq 2T} \left| \frac{n}{k} \tilde{F}_{n, \alpha, \beta} \left(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z} \right) - \frac{n}{k} \tilde{F}_{\alpha, \beta} \left(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z} \right) \right|.$$

Thus by Lemma 9.21, with probability at least $1 - 2\delta$,

$$\Lambda(n) \leq Cd \sqrt{\frac{2T}{k} \log \frac{1}{\delta}}.$$

Concerning $\Upsilon(n)$, we have the following decomposition:

$$\begin{aligned}
\Upsilon(n) &\leq \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} \left| g_{\alpha, \beta} \left(\frac{n}{k} \left(U_{(\lfloor kx_j \rfloor)}^j \right)_{j \in \alpha}, \frac{n}{k} \left(U_{(\lfloor kz_j \rfloor)}^j \right)_{j \in \beta} \right) \right. \\
&\quad \left. - g_{\alpha, \beta} \left(\left(\frac{\lfloor kx_j \rfloor}{k} \right)_{j \in \alpha}, \left(\frac{\lfloor kz_j \rfloor}{k} \right)_{j \in \beta} \right) \right| \\
&\quad + \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} \left| g_{\alpha, \beta} \left(\left(\frac{\lfloor kx_j \rfloor}{k} \right)_{j \in \alpha}, \left(\frac{\lfloor kz_j \rfloor}{k} \right)_{j \in \beta} \right) - g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) \right| \\
&=: \Upsilon_1(n) + \Upsilon_2(n).
\end{aligned}$$

The inequality in Lemma 9.6 allows us to bound the first term $\Upsilon_1(n)$:

$$\begin{aligned}
\Upsilon_1(n) &\leq C \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} \sum_{j \in \alpha} \left| \frac{\lfloor kx_j \rfloor}{k} - \frac{n}{k} U_{(\lfloor kx_j \rfloor)}^j \right| + \sum_{j \in \beta} \left| \frac{\lfloor kz_j \rfloor}{k} - \frac{n}{k} U_{(\lfloor kz_j \rfloor)}^j \right| \\
&\leq 2C \sup_{0 \leq \mathbf{x} \leq T} \sum_{1 \leq j \leq d} \left| \frac{\lfloor kx_j \rfloor}{k} - \frac{n}{k} U_{(\lfloor kx_j \rfloor)}^j \right|
\end{aligned}$$

so that by Lemma 9.23, with probability greater than $1 - (d + 1)\delta$:

$$\Upsilon_1(n) \leq Cd \sqrt{\frac{2T}{k} \log \frac{1}{\delta}}.$$

Similarly,

$$\Upsilon_2(n) \leq 2C \sup_{0 \leq \mathbf{x} \leq T} \sum_{1 \leq j \leq d} \left| \frac{\lfloor kx_j \rfloor}{k} - x_j \right| \leq C \frac{2d}{k}.$$

Finally we get, for every $n > 0$, with probability at least $1 - (d + 3)\delta$,

$$\begin{aligned} \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq T} |g_{n, \alpha, \beta}(\mathbf{x}, \mathbf{z}) - g_{\alpha, \beta}(\mathbf{x}, \mathbf{z})| &\leq \Lambda(n) + \Upsilon_1(n) + \Upsilon_2(n) + \Xi(n) \\ &\leq Cd \sqrt{\frac{2T}{k} \log \frac{1}{\delta}} + \frac{2d}{k} + \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq 2T} \left| \frac{n}{k} \tilde{F}_{\alpha, \beta}(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z}) - g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) \right| \\ &\leq C'd \sqrt{\frac{2T}{k} \log \frac{1}{\delta}} + \max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq 2T} \left| \frac{n}{k} \tilde{F}_{\alpha, \beta}(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z}) - g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) \right|. \end{aligned}$$

Remark 9.24. (BIAS TERM) It is classical (see Qi (1997) p.174 for details) to extend the simple convergence (9.27) to the uniform version on $[0, T]^d$. It suffices to subdivide $[0, T]^d$ and to use the monotonicity in each dimension coordinate of $g_{\alpha, \beta}$ and $\tilde{F}_{\alpha, \beta}$. Thus,

$$\sup_{0 \leq \mathbf{x}, \mathbf{z} \leq 2T} \left| \frac{n}{k} \tilde{F}_{\alpha, \beta}(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z}) - g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) \right| \rightarrow 0$$

for every α and β . Note also that by taking a maximum on a finite class we have the convergence of the maximum uniform bias to 0:

$$\max_{\alpha, \beta} \sup_{0 \leq \mathbf{x}, \mathbf{z} \leq 2T} \left| \frac{n}{k} \tilde{F}_{\alpha, \beta}(\frac{k}{n} \mathbf{x}, \frac{k}{n} \mathbf{z}) - g_{\alpha, \beta}(\mathbf{x}, \mathbf{z}) \right| \rightarrow 0. \quad (9.48)$$

9.7.4 Proof of Lemma 9.10

First note that as the Ω_β 's form a partition of the simplex S_∞^{d-1} and that $\Omega_\alpha^{\epsilon, \epsilon'} \cap \Omega_\beta = \emptyset$ as soon as $\alpha \not\subset \beta$, we have

$$\Omega_\alpha^{\epsilon, \epsilon'} = \bigsqcup_{\beta} \Omega_\alpha^{\epsilon, \epsilon'} \cap \Omega_\beta = \bigsqcup_{\beta \supset \alpha} \Omega_\alpha^{\epsilon, \epsilon'} \cap \Omega_\beta.$$

Let us recall that as stated in Lemma 9.3), Φ is concentrated on the (disjoint) edges

$$\Omega_{\alpha, i_0} = \{ \mathbf{x} : \|\mathbf{x}\|_\infty = 1, x_{i_0} = 1, 0 < x_i < 1 \text{ for } i \in \alpha \setminus \{i_0\} \\ x_i = 0 \text{ for } i \notin \alpha \}$$

and that the restriction Φ_{α, i_0} of Φ to Ω_{α, i_0} is absolutely continuous w.r.t. the Lebesgue measure $dx_{\alpha \setminus i_0}$ on the cube's edges, whenever $|\alpha| \geq 2$. By (9.15) we have, for every $\beta \supset \alpha$,

$$\begin{aligned} \Phi(\Omega_\alpha^{\epsilon, \epsilon'} \cap \Omega_\beta) &= \sum_{i_0 \in \beta} \int_{\Omega_\alpha^{\epsilon, \epsilon'} \cap \Omega_{\beta, i_0}} \frac{d\Phi_{\beta, i_0}}{dx_{\beta \setminus i_0}}(x) dx_{\beta \setminus i_0} \\ \Phi(\Omega_\alpha) &= \sum_{i_0 \in \alpha} \int_{\Omega_{\alpha, i_0}} \frac{d\Phi_{\alpha, i_0}}{dx_{\alpha \setminus i_0}}(x) dx_{\alpha \setminus i_0}. \end{aligned}$$

Thus,

$$\begin{aligned}
\Phi(\Omega_{\alpha}^{\epsilon, \epsilon'}) - \Phi(\Omega_{\alpha}) &= \sum_{\beta \supset \alpha} \sum_{i_0 \in \beta} \int_{\Omega_{\alpha}^{\epsilon, \epsilon'} \cap \Omega_{\beta, i_0}} \frac{d\Phi_{\beta, i_0}}{dx_{\beta \setminus i_0}}(x) dx_{\beta \setminus i_0} \\
&\quad - \sum_{i_0 \in \alpha} \int_{\Omega_{\alpha, i_0}} \frac{d\Phi_{\alpha, i_0}}{dx_{\alpha \setminus i_0}}(x) dx_{\alpha \setminus i_0} \\
&= \sum_{\beta \supset \alpha} \sum_{i_0 \in \beta} \int_{\Omega_{\alpha}^{\epsilon, \epsilon'} \cap \Omega_{\beta, i_0}} \frac{d\Phi_{\beta, i_0}}{dx_{\beta \setminus i_0}}(x) dx_{\beta \setminus i_0} \\
&\quad - \sum_{i_0 \in \alpha} \int_{\Omega_{\alpha, i_0} \setminus (\Omega_{\alpha}^{\epsilon, \epsilon'} \cap \Omega_{\alpha, i_0})} \frac{d\Phi_{\alpha, i_0}}{dx_{\alpha \setminus i_0}}(x) dx_{\alpha \setminus i_0},
\end{aligned}$$

so that by eq9.16,

$$\begin{aligned}
|\Phi(\Omega_{\alpha}^{\epsilon, \epsilon'}) - \Phi(\Omega_{\alpha})| &\leq \sum_{\beta \supset \alpha} M_{\beta} \sum_{i_0 \in \beta} \int_{\Omega_{\alpha}^{\epsilon, \epsilon'} \cap \Omega_{\beta, i_0}} dx_{\beta \setminus i_0} \\
&\quad + M_{\alpha} \sum_{i_0 \in \alpha} \int_{\Omega_{\alpha, i_0} \setminus (\Omega_{\alpha}^{\epsilon, \epsilon'} \cap \Omega_{\alpha, i_0})} dx_{\alpha \setminus i_0}.
\end{aligned} \tag{9.49}$$

Without loss of generality we may assume that $\alpha = \{1, \dots, K\}$ with $K \leq d$. Then, for $\beta \supset \alpha$, $\int_{\Omega_{\alpha}^{\epsilon, \epsilon'} \cap \Omega_{\beta, i_0}} dx_{\beta \setminus i_0}$ is smaller than $(\epsilon')^{|\beta| - |\alpha|}$ and is null as soon as $i_0 \in \beta \setminus \alpha$. To see this, assume for instance that $\beta = \{1, \dots, P\}$ with $P > K$. Then

$$\begin{aligned}
\Omega_{\alpha}^{\epsilon, \epsilon'} \cap \Omega_{\beta, i_0} &= \{\epsilon < x_1, \dots, x_K \leq 1, x_{K+1}, \dots, x_P \leq \epsilon', x_{i_0} = 1, \\
&\quad x_{P+1} = \dots = x_d = 0\}
\end{aligned}$$

which is empty if $i_0 \geq K + 1$ (i.e. $i_0 \in \beta \setminus \alpha$) and which fulfills if $i_0 \leq K$

$$\int_{\Omega_{\alpha}^{\epsilon, \epsilon'} \cap \Omega_{\beta, i_0}} dx_{\beta \setminus i_0} \leq (\epsilon')^{P-K}.$$

The first term in (9.49) is then bounded by $\sum_{\beta \supset \alpha} M_{\beta} |\alpha| (\epsilon')^{|\beta| - |\alpha|}$. Now, concerning the second term in (9.49), $\Omega_{\alpha}^{\epsilon, \epsilon'} \cap \Omega_{\alpha, i_0} = \{\epsilon < x_1, \dots, x_K \leq 1, x_{i_0} = 1, x_{K+1}, \dots, x_d = 0\}$ and then

$$\Omega_{\alpha, i_0} \setminus (\Omega_{\alpha}^{\epsilon, \epsilon'} \cap \Omega_{\alpha, i_0}) = \bigcup_{l=1, \dots, K} \Omega_{\alpha, i_0} \cap \{x_l \leq \epsilon\},$$

so that $\int_{\Omega_{\alpha, i_0} \setminus (\Omega_{\alpha}^{\epsilon, \epsilon'} \cap \Omega_{\alpha, i_0})} dx_{\alpha \setminus i_0} \leq K\epsilon = |\alpha|\epsilon$. The second term in (9.49) is thus bounded by $M|\alpha|^2\epsilon$. Finally, (9.49) implies

$$|\Phi(\Omega_{\alpha}^{\epsilon, \epsilon'}) - \Phi(\Omega_{\alpha})| \leq |\alpha| \sum_{\beta \supset \alpha} M_{\beta} (\epsilon')^{|\beta| - |\alpha|} + M|\alpha|^2\epsilon.$$

To conclude, observe that by Assumption 3,

$$\sum_{\beta \supset \alpha} M_{\beta} (\epsilon')^{|\beta| - |\alpha|} \leq \sum_{\beta \supset \alpha} M_{\beta} (\epsilon') \leq \epsilon' \sum_{|\beta| \geq 2} M_{\beta} \leq \epsilon' M$$

The result is thus proved.

9.7.5 Proof of Remark 9.14

Let us prove that Z_n , conditionally to the event $\{\|\frac{k}{n}\mathbf{V}_1\|_\infty \geq 1\}$, converges in law. Recall that Z_n is a $(2^d - 1)$ -vector defined by $Z_n(\alpha) = \mathbb{1}_{\frac{k}{n}\mathbf{V}_1 \in R_\alpha^\epsilon}$ for all $\alpha \subset \{1, \dots, d\}, \alpha \neq \emptyset$. Let us denote $1_\alpha = (\mathbb{1}_{j=\alpha})_{j=1, \dots, 2^d-1}$ where we implicitly define the bijection between $\mathcal{P}(\{1, \dots, d\}) \setminus \emptyset$ and $\{1, \dots, 2^d - 1\}$. Since the R_α^ϵ 's, α varying, form a partition of $[\mathbf{0}, \mathbf{1}]^c$, $\mathbb{P}(\exists \alpha, Z_n = 1_\alpha \mid \|\frac{k}{n}\mathbf{V}_1\|_\infty \geq 1) = 1$ and $Z_n = 1_\alpha \Leftrightarrow Z_n(\alpha) = 1 \Leftrightarrow \frac{k}{n}\mathbf{V}_1 \in R_\alpha^\epsilon$, so that

$$\mathbb{E} \left[\Phi(Z_n) \mathbb{1}_{\|\frac{k}{n}\mathbf{V}_1\|_\infty \geq 1} \right] = \sum_{\alpha} \Phi(1_\alpha) \mathbb{P}(Z_n(\alpha) = 1).$$

Let $\Phi : \mathbb{R}^{2^d-1} \rightarrow \mathbb{R}_+$ be a measurable function. Then

$$\mathbb{E} \left[\Phi(Z_n) \mid \|\frac{k}{n}\mathbf{V}_1\|_\infty \geq 1 \right] = \mathbb{P} \left[\|\frac{k}{n}\mathbf{V}_1\|_\infty \geq 1 \right]^{-1} \mathbb{E} \left[\Phi(Z_n) \mathbb{1}_{\|\frac{k}{n}\mathbf{V}_1\|_\infty \geq 1} \right].$$

Now, $\mathbb{P} \left[\|\frac{k}{n}\mathbf{V}_1\|_\infty \geq 1 \right] = \frac{k}{n} \pi_n$ with $\pi_n \rightarrow \mu([\mathbf{0}, \mathbf{1}]^c)$, so that

$$\mathbb{E} \left[\Phi(Z_n) \mid \|\frac{k}{n}\mathbf{V}_1\|_\infty \geq 1 \right] = \pi_n^{-1} \frac{n}{k} \left(\sum_{\alpha} \Phi(1_\alpha) \mathbb{P}(Z_n(\alpha) = 1) \right).$$

Using $\frac{n}{k} \mathbb{P}[Z_n(\alpha) = 1] = \frac{n}{k} \mathbb{P}[\frac{k}{n}\mathbf{V}_1 \in R_\alpha^\epsilon] \rightarrow \mu(R_\alpha^\epsilon)$, we find that

$$\mathbb{E} \left[\Phi(Z_n) \mid \|\frac{k}{n}\mathbf{V}_1\|_\infty \geq 1 \right] \rightarrow \sum_{\alpha} \Phi(1_\alpha) \frac{\mu(R_\alpha^\epsilon)}{\mu([\mathbf{0}, \mathbf{1}]^c)},$$

which achieves the proof.

9.8 Experiments curves

CHAPTER 10

Conclusion & Perspectives

In this thesis, we have addressed three important limitations of existing anomaly detection literature. The problem of evaluating algorithm in the case of unlabeled data, the problem of extending the use of random forests to one-class classification, and the problem of being accurate on low probability regions.

Our first contribution was to proposed an unsupervised performance criterion, in order to compare scoring functions and to pick one eventually. This excess-mass based criterion resolved some of the drawbacks inherent to the previous mass-volume curve criterion. But the main drawback, estimating a volume in the input space, still remains. This unfortunately constitutes a strong setback for its use in a high dimensional framework, if no prior knowledge on the form of these volume to estimate is available. In such a context, classical evaluation approaches must be used, such as ROC or Precision-Recall curves, which assume that at least a small proportion of data is labelled. Practical unsupervised evaluation criteria still remaining a major challenge (specially under the constraint of scaling with dimension), we studied empirically the use of EM or MV as evaluation criteria and proposed a way to scale their use to high dimensions.

As trying to minimize EM or MV criteria does not produce performant algorithms in practice, we introduced a One Class Random Forest algorithm which structurally extend RFs to one-class classification.

Finally, we proposed to focus on extreme regions to gain in accuracy when building scoring functions. An intermediary step was to studies non-asymptotic behavior of an extreme value copula, the stable tail dependence function. We brought new bounds to control the error of its natural empirical version as well as a practical framework for deriving VC-type bounds on low-probability regions. This framework also allows to approach a particular prediction context, namely where the objective is to learn a classifier (or a regressor) that has good properties on low-probability regions. Besides, as we exhibit a sparsity pattern in multivariate extremes, it can be used as a preprocessing step to scale up multivariate extreme values modeling to high dimensional settings, which is currently one of the major challenges in multivariate EVT.

Staying in the scope of multivariate extremes, the non-asymptotic bounds in the two main results contain separated bias terms corresponding to the (distribution-dependent) convergence speed to the asymptotic behavior, which are not controlled explicitly. A possible future direction is to

make an additional hypothesis of ‘second order regular variation’ (see *e.g.* de Haan & Resnick, 1996) in order to express these bias terms, and possibly to refine the results. With such explicit bounds, parameters of the Damex algorithm (third contribution) could be chosen optimally as the ones minimizing the obtained bound.

From the scope of one-class random forests, a possible research direction would be to develop theoretical grounds for the level sets estimation procedure. Classical studies from the two-class framework should be adapted to one-class classification.

Bibliography

- C.C. Aggarwal and P.S. Yu. Outlier detection for high dimensional data. In *ACM Sigmod Record*, 2001.
- Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural comp.*, 1997.
- M. Anthony and J. Shawe-Taylor. A result of vapnik with applications. *Discrete Applied Mathematics*, 47(3):207 – 217, 1993.
- V. Barnett and T. Lewis. *Outliers in statistical data*, volume 3. Wiley New York, 1994.
- J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. Wiley, 2004.
- J. Beirlant, P. Vynckier, and J. L. Teugels. Tail index estimation, pareto quantile plots regression diagnostics. *Journal of the American Statistical Association*, 91(436):1659–1667, 1996.
- Jan Beirlant, Mikael Escobar-Bach, Yuri Goegebeur, and Armelle Guillou. Bias-corrected estimation of stable tail dependence function. <https://hal.archives-ouvertes.fr/hal-01115538>, Feb 2015.
- G rard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.
- Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 2010.
- L on Bottou and Chih-Jen Lin. Support vector machine solvers. *Large scale kernel machines*, pages 301–320, 2007.
- S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
- O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar R tsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004.
- L. Breiman. Random forests. *Machine Learning*, 2001. ISSN 0885-6125.

- M.M. Breunig, H.P. Kriegel, R.T. Ng, and J. Sander. LOF: identifying density-based local outliers. In *ACM sigmod record*, 2000.
- Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.
- V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 2009.
- S. Cléménçon and J. Jakubowicz. Scoring anomalies: a M-estimation approach. In *AISTATS*, 2013.
- S. Cléménçon and S. Robbiano. Anomaly Ranking as Supervised Bipartite Ranking. In *ICML*, 2014.
- S. Cléménçon and N. Vayatis. Nonparametric estimation of the precision-recall curve. In *ICML*, 2009.
- Stéphan Cléménçon and Nicolas Vayatis. Overlaying classifiers: a practical approach to optimal scoring. *Constructive Approximation*, 32(3):619–648, 2010.
- D.A. Clifton, L. Tarassenko, N. McGrogan, D. King, S. King, and P. Anuzis. Bayesian extreme value statistics for novelty detection in gas-turbine engines. In *Aerospace Conference, 2008 IEEE*, pages 1–11, 2008.
- David Andrew Clifton, Samuel Hugueny, and Lionel Tarassenko. Novelty detection with multivariate extreme value statistics. *Journal of signal processing systems*, 65(3):371–389, 2011.
- S. Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag, London, 2001.
- SG Coles and JA Tawn. Modeling extreme multivariate events. *JR Statist. Soc. B*, 53:377–392, 1991.
- D. Cooley, R.A. Davis, and P. Naveau. The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117, 2010.
- J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *ICML*, 2006.
- L. de Haan and A. Ferreira. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, 2006. An introduction.
- L. de Haan and S. Resnick. Second-order regular variation and rates of convergence in extreme-value theory. *The Annals of Probability*, pages 97–124, 1996.

- L. de Haan and S.I. Resnick. Limit theory for multivariate sample extremes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 40(4):317–337, 1977.
- A. L. M. Dekkers, J. H. J. Einmahl, and L. de Haan. A moment estimator for the index of an extreme-value distribution. *Ann. Statist.*, 17(4):1833–1855, 12 1989.
- L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of mathematics : stochastic modelling and applied probability. U.S. Government Printing Office, 1996.
- H. Drees and X. Huang. Best attainable rates of convergence for estimators of the stable tail dependence function. *J. Multivar. Anal.*, 64(1):25–47, January 1998.
- C. Désir, S. Bernard, C. Petitjean, and L. Heutte. A new random forest method for one-class classification. In *Structural, Syntactic, and Statistical Pattern Recognition*. Springer, 2012.
- J. H. J. Einmahl, L. de Haan, and D. Li. Weighted approximations of tail copula processes with application to testing the bivariate extreme value condition. *Ann. Statist.*, 34(4):1987–2014, 08 2006.
- J. H. J. Einmahl, A. Krajina, and J. Segers. An m-estimator for tail dependence in arbitrary dimensions. *Ann. Statist.*, 40:1764–1793, 2012.
- J. H. J. Einmahl, J. Li, and R. Y. Liu. Thresholding events of extreme in simultaneous monitoring of multiple risks. *Journal of the American Statistical Association*, 104(487):982–992, 2009.
- J. H. J. Einmahl and J. Segers. Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, pages 2953–2989, 2009.
- John HJ Einmahl, Laurens de Haan, and Vladimir I Piterbarg. Nonparametric estimation of the spectral measure of an extreme value distribution. *Annals of Statistics*, pages 1401–1423, 2001.
- John HJ Einmahl and David M Mason. Generalized quantile processes. *The Annals of Statistics*, pages 1062–1078, 1992.
- P. Embrechts, L. de Haan, and X. Huang. Modelling multivariate extremes. *Extremes and Integrated Risk Management (Ed. P. Embrechts)*, RISK Books(59-67), 2000.
- E. Eskin. Anomaly detection over noisy data using learned probability distributions. In *ICML*, 2000.
- E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pages 77–101. Springer, 2002.
- M. Falk, J. Huesler, and R. D. Reiss. *Laws of Small Numbers: Extremes and Rare Events*. Birkhauser, Boston, 1994.

- T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- H. Federer. *Geometric Measure Theory*. Springer, 1969.
- Barbel Finkenstadt and Holger Rootzén. *Extreme values in finance, telecommunications, and the environment*. CRC Press, 2003.
- A.-L. Fougères, L. De Haan, and C. Mercadier. Bias correction in multivariate extremes. *The Annals of Statistics*, 43(2):903–934, 2015.
- Anne-Laure Fougères, John P Nolan, and Holger Rootzén. Models for dependent extremes using stable mixtures. *Scandinavian Journal of Statistics*, 36(1):42–59, 2009.
- R. Genuer, J.-M. Poggi, and C. Tuleau. Random forests: some methodological insights. *arXiv:0811.3619*, 2008.
- C. Gini. Variabilità e mutabilità. *Memorie di metodologia statistica*, 1912.
- N. Goix. How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms? In *ICML Workshop on Anomaly Detection*, 2016.
- N. Goix, R. Brault, N. Drougard, and M. Chiapino. One Class Splitting Criteria for Random Forests with Application to Anomaly Detection. In *Submitted to NIPS*, 2016a.
- N. Goix, A. Sabourin, and S. Cléménçon. Sparse Representation of Multivariate Extremes. NIPS 2015 Workshop on Nonparametric Methods for Large Scale Representation Learning, 2015a.
- N. Goix, A. Sabourin, and S. Cléménçon. Learning the dependence structure of rare events: a non-asymptotic study. In *COLT*, 2015b.
- N. Goix, A. Sabourin, and S. Cléménçon. On Anomaly Ranking and Excess-Mass Curves. In *AISTATS*, 2015c.
- N. Goix, A. Sabourin, and S. Cléménçon. Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. In *AISTAT*, 2016b.
- N. Goix and A. Thomas. How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms? In *Submitted to NIPS*, 2016.
- Nicolas Goix, Anne Sabourin, and Stéphan Cléménçon. Sparse Representation of Multivariate Extremes with Applications to Anomaly Detection. Submitted to Journal of Multivariate Analysis, July 2016c.
- J.A. Hartigan. Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, 82(397):267–270, 1987.
- B. M. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3(5):1163–1174, 09 1975.

- T.K. Ho. The random subspace method for constructing decision forests. *TPAMI*, 1998.
- V.J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artif. Intel. Review*, 2004.
- X. Huang. Statistics of bivariate extreme values, 1992.
- Svante Janson. On concentration of probability. *Contemporary combinatorics*, 2002.
- Eric Jones, Travis Oliphant, Pearu Peterson, et al. Scipy: Open source scientific tools for python, 2001–. URL <http://www.scipy.org>, 2015.
- KDDCup. The third international knowledge discovery and data mining tools competition dataset. 1999.
- V. Koltchinskii. M-estimation, convexity and quantiles. *The Annals of Statistics*, 25:435–477, 1997.
- V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization (with discussion). *The Annals of Statistics*, 34:2593–2706, 2006.
- M Ross Leadbetter, Georg Lindgren, and Holger Rootzén. Extremes and related properties of random sequences and processes. *Springer Series in Statistics*, 1983.
- H.J. Lee and S.J. Roberts. On-line novelty detection using the kalman filter and extreme value theory. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4, 2008.
- M. Lichman. UCI machine learning repository, 2013.
- R. Lippmann, J. W Haines, D.J. Fried, J. Korba, and K. Das. Analysis and results of the 1999 darpa off-line intrusion detection evaluation. In *Recent Advances in Intrusion Detection*, pages 162–182. Springer, 2000.
- F.T. Liu, K.M. Ting, and Z.H. Zhou. Isolation Forest. In *ICDM*, 2008.
- M. Markou and S. Singh. Novelty detection: a review part 1: statistical approaches. *Signal proc.*, 2003.
- Pascal Massart. Some applications of concentration inequalities to statistics. *Annales de la faculté des sciences de Toulouse*, 9(2):245–303, 2000.
- Colin McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, Algorithms and Combinatorics. Springer, 1998.
- D.W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415):738–746, 1991.

- P. Panov and S. Džeroski. *Combining bagging and random subspaces to create better ensembles*. Springer, 2007.
- A. Patcha and J.M. Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 2007.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *JMLR*, 2011.
- W. Polonik. Measuring Mass Concentrations and Estimating Density Contour Cluster-An excess Mass Approach. *The Annals of Statistics*, 1995.
- W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 1997.
- W. Polonik. The silhouette, concentration functions and ml-density estimation under order restrictions. *The Annals of Statistics*, 26, 1998.
- FJ Provost, T. Fawcett, et al. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In *KDD*, volume 97, pages 43–48, 1997.
- FJ Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *ICML*, 1998.
- Y. Qi. Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics. *Acta Mathematicae Applicatae Sinica*, 13(2):167–175, 1997.
- S. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Springer Series in Operations Research and Financial Engineering, 1987.
- Sidney Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.
- S.J. Roberts. Novelty detection using extreme value statistics. *Vision, Image and Signal Processing, IEE Proceedings -*, 146(3):124–129, Jun 1999.
- S.J. Roberts. Extreme value statistics for novelty detection in biomedical signal processing. In *Advances in Medical Signal and Information Processing, 2000. First International Conference on (IEE Conf. Publ. No. 476)*, pages 166–172, 2000.
- Anne Sabourin and Philippe Naveau. Bayesian dirichlet mixture model for multivariate extremes: A re-parametrization. *Computational Statistics & Data Analysis*, 2012.
- B. Schölkopf, J.C Platt, J. Shawe-Taylor, A.J Smola, and R.C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.
- C.D Scott and R.D Nowak. Learning minimum volume sets. *The Journal of Machine Learning Research*, 7:665–704, 2006.

- J. Segers. Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, 18(3):764–782, 08 2012a.
- J. Segers. Max-stable models for multivariate extremes. *REVSTAT - Statistical Journal*, 10(1): 61–82, 2012b.
- T. Shi and S. Horvath. Unsupervised learning with random forest predictors. *J. Comp. Graph. Stat.*, 2012.
- M.L. Shyu, S.C. Chen, K. Sarinnapakorn, and L. Chang. A novel anomaly detection scheme based on principal component classifier. Technical report, DTIC Document, 2003.
- R. L. Smith. Estimating tails of probability distributions. *Ann. Statist.*, 15(3):1174–1207, 09 1987.
- RL Smith. Statistics of extremes, with applications in environment, insurance and finance, chap 1. *Statistical analysis of extreme values: with applications to insurance, finance, hydrology, and other fields*. Birkhäuser, Basel, 2003.
- I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005.
- A.G. Stephenson. High-dimensional parametric modelling of multivariate extreme events. *Australian & New Zealand Journal of Statistics*, 51(1):77–88, 2009.
- Alec Stephenson. Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6(1):49–59, 2003.
- M. Tavallaei, E. Bagheri, W. Lu, and A.A. Ghorbani. A detailed analysis of the kdd cup 99 data set. In *IEEE CISDA*, 2009.
- JA Tawn. Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245–253, 1990.
- A. Thomas, V. Feuillard, and A. Gramfort. Calibration of One-Class SVM for MV set estimation. In *DSAA*, 2015.
- Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 2011.
- V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, 1974. (German Translation: W. Wapnik & A. Tscherwonienkis, *Theorie der Zeichenerkennung*, Akademie-Verlag, Berlin, 1979).
- J.P. Vert and R. Vert. Consistency and convergence rates of one-class svms and related algorithms. *JMLR*, 6:828–835, 2006.
- K. Viswanathan, L. Choudur, V. Talwar, C. Wang, G. Macdonald, and W. Satterfield. Ranking anomalies in data centers. In R.D.James, editor, *Network Operations and System Management*, pages 79–87. IEEE, 2012.

- JA. Wellner. Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 45(1): 73–88, 1978.
- K. Yamanishi, J.I. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *KDD*, 2000.

Abstract

Résumé