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 28 Novembre 2016

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

Jury :

Gérard **Biau**	Professeur, Université Pierre et Marie Curie	Examinateur
Stéphane **Boucheron**	Professeur, Université Paris Diderot	Rapporteur
Stéphan **Clémençon**	Professeur, Télécom ParisTech	Directeur
Stéphane **Girard**	Directeur de Recherche, Inria Grenoble Rhône-Alpes	Rapporteur
Alexandre **Gramfort**	Maitre de Conférence, Télécom ParisTech	Examinateur
Anne **Sabourin**	Maitre de Conférence, Télécom ParisTech	Co-directeur
Jean-Philippe **Vert**	Directeur de Recherche, Mines ParisTech	Examinateur

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 Clémençon.

Conferences

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

Je souhaite tout d'abord remercier mes directeurs de thèse, Stéphan et Anne; la complémentarité de vos originalités respectives constitue un magma favorable à la recherche. En particulier, merci Stéphan pour l'originalité du sujet de thèse et pour ton recul sur la littérature. Merci aussi pour le dynamisme apporté à l'équipe STA, qui s'est créée une réputation internationale en marquant fièrement sa présence aux grosses conférences de machine learning.

Contents

List of Figures

List of Tables

Dans ce prodigieux, cet inépuisable catalogue-raisonné-des-données-expérimentales, qui, mené à bien, devait aboutir à la certitude des certitudes, toutes les bizarreries, toutes les anomalies avaient leur place marquée, puisqu'elles étaient révélatrices à l'égal des phénomènes " normaux ". BÉGUIN, L'Âme romantique et le rêve, 1939, p. 4.

CHAPTER 1 Summary

1.1 Introduction

Anomaly, from Greek $\alpha\nu\omega\mu\alpha\lambda\iota\alpha$, 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 a fixed pathology or not, using previous patients record (such as age, history, gender, blood pressure) associated with their true **label** (having the pathology or not). This case is an example of *binary classification*, where 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 accurately.

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 **novelty detection** framework (also called one-class classification or semi-supervised anomaly 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. Section 1.4 focuses on extreme value theory (EVT) to gain in accuracy on extreme regions. It introduces the *stable tail dependence function* (STDF), to estimate the dependence structure of rare events (Section 1.4.1) and shows that multivariate EVT can be useful to produce scoring functions accurate on low probability regions (Section 1.4.2). Section 1.5 gathers two contributions to be submitted addressing the evaluation of unsupervised anomaly detection from a practical point of view (Section 1.5.1), and the extension of random forests to the one-class setting (Section 1.5.2). Section 1.6 presents contributions relative to a widely used open-source machine learning library. Section 1.7 details the scientific output and concludes.

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 non-negative 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. $X \overset{d}{=} Y$ means that X and Y are equal in distribution and $X_n \overset{d}{\to} Y$ 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, Anomaly Ranking and Scoring Functions

From a probabilistic point of view, there are different ways 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 $(\mathrm{d}F/\mathrm{d}G)(\mathbf{x}) > t$ 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* $\{\mathbf{x}, (\mathrm{d}F/\mathrm{d}G)(\mathbf{x}) > t\}$ (Schölkopf et al., 2001; Steinwart et al., 2005; Scott & Nowak, 2006; Vert & Vert, 2006). 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 (Schölkopf et al., 2001; Scott & Nowak, 2006; Vert & Vert, 2006).

This comes back to assume that anomalies are uniformly distributed on the support of the normal distribution. 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 $f : \mathbb{R}^d \to \mathbb{R}$. 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.

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 sets (Tsybakov, 1997; Cuevas & Fraiman,

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
$\lvert\mathcal{G}\rvert$	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 $a \in \mathbb{R}^d$
$\lfloor \cdot \rfloor$	integer part
$\langle \cdot, \cdot \rangle$	Inner product between vectors
$\lVert \cdot \rVert$	An arbitrary norm
$\lVert \cdot \rVert_p$	L_p norm
$A \Delta B$	symmetric difference between sets A and B
$(\Omega, \mathcal{F}, \mathbb{P})$	Underlying probability space
$\mathcal{S}$	functions $s : \mathbb{R}^d \to \mathbb{R}_+$ 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 $\mathcal{E}$
$Y_{(1)} \leq \ldots \leq Y_{(n)}$	order statistics of $Y_1, \dots, Y_n$
$\mathbf{X}_{1:n}$	An $i.i.d.$ sample $(\mathbf{X}_1, \dots, \mathbf{X}_n)$
$\mathbb{P}[\cdot]$	Probability of event
$\mathbb{E}[\cdot]$	Expectation of random variable
$\mathbf{Var}[\cdot]$	Variance of random variable

TABLE 1.1: Summary of notation.

1997; Baillo et al., 2001; Baillo, 2003; Cadre, 2006; Rigollet & Vert, 2009; Mason & Polonik, 2009). 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. They try somehow to capture more information than needed for the 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 representative of the class of all increasing transforms of f, to obtain density level sets estimates. Intuitively, it is enough to estimate the pre-order (the *scoring*) induced by f on $\mathbb{R}^d$. Let us define a *scoring function* as any measurable function $s : \mathbb{R}^d \to \mathbb{R}_+$ integrable *w.r.t.* the Lebesgue measure Leb(.), and $\mathcal{S}$ the space of all scoring functions. Any scoring function defines a pre-order 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 return a scoring function, which can be 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 pre-order 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 $\Phi : \mathbb{R} \to \mathbb{R}_+$ 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.

Remark 1.1. **Terminology: anomaly detection, anomaly ranking.**
Strictly speaking, criteria we are looking for are anomaly *ranking* criteria, in the same way the ROC curve is basically a criterion for bipartite *ranking*. In practice as mentioned above, all the algorithms dealing with anomalies are candidates for the anomaly ranking task. They all produce a ranking/scoring function, even the ones which originally deal with the 'anomaly classification' framework, *i.e.* seek to be optimal on a single level set or *w.r.t.* a fixed false positive rate. In the literature, the 'anomaly detection' terminology is widely used, instead of the more precise 'anomaly ranking' one. For instance, Liu et al. (2008) write '*The task of anomaly detection is to provide a ranking that reflects the degree of anomaly*'. In this work, we similarly make the convention that anomaly detection refers to anomaly ranking: if labels are available for the evaluation step, the goal is to maximize the area under the ROC curve. If no labeled data is available, the goal is to maximize the alternative criteria defined in the next section.

1.3 M-estimation and criteria for scoring functions

This section is a summary of Chapter 6, 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 Clémençon & 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

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

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 remaining 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:

$\mathbf{A_1}$ *The density f is bounded.*

$\mathbf{A_2}$ *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:

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

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 $\widehat{\Gamma}^*_\alpha$ 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 $\{\text{Leb}(\widehat{\Gamma}^*_\alpha) - \lambda^*(\alpha)\}$ (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 Clémençon & Jakubowicz (2013); Clémençon & 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 $H : \mathbb{R} \to (0,1)$ and where α_s and λ_s are defined by

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

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

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

Under assumptions $\mathbf{A_1}$ and $\mathbf{A_2}$, 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 $\mathbf{X}_1, ...\mathbf{X}_n$ such that $MV_{\hat{s}}$ is (nearly) minimum everywhere, *i.e.* minimizing $\|MV_{\hat{s}} - MV^*\|_\infty := \sup_{\alpha\in[0,1]} |MV_{\hat{s}}(\alpha) - MV^*(\alpha)|$.

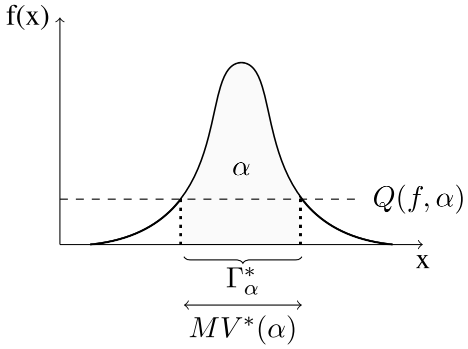


FIGURE 1.1: Mass-Volume at level α

The way of doing it consists in preliminarily estimating a collection of minimum volume sets related to target masses $0 < \alpha_1 < \ldots < \alpha_K < 1$ forming a subdivision of $(0,1)$ based on training data so as to define $s = \sum_k \mathbb{1}_{\{x \in \hat{\Gamma}^*_{\alpha_k}\}}$. 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 pre-orders/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 $t > 0$,

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

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

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

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 $\mathbf{A_1}$, we have $EM_s(t) = 0$ for every $t > \|f\|_\infty$.

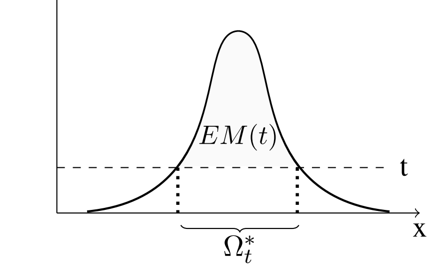


Figure 2: Excess-Mass curve

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, $EM_s(t) \le EM^*(t)$ for all $t \ge 0$. In addition, for $t \ge 0$ and any $\epsilon > 0$, we have

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

with $\{s > u\} \Delta_\epsilon \{f > t\} \;:=\; \{f > t + \epsilon\} \setminus \{s > u\} \;\bigsqcup\; \{s > u\} \setminus \{f > t - \epsilon\}$. 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 $\epsilon > 0$,

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

where the infimum is taken over the set $\mathcal{T}$ of all measurable increasing transforms $T : \mathbb{R}_+ \to \mathbb{R}_+$. 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 pre-orders 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 Clémençon & 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

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

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 $\{s \ge t_k\}$. Then, maximizing the Excess-Mass functional 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 6.

1.4 Accuracy on extreme regions

1.4.1 Extreme Values Analysis through STDF estimation

This section is a summary of Chapter 7, 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 $x_p = \inf\{x \in \mathbb{R},\ \mathbb{P}(X > x) \le p\}$. For moderate values of p, a natural empirical estimate is $x_{p,n} = \inf\{x \in \mathbb{R},\ 1/n \sum_{i=1}^n \mathbb{1}_{X_i > x} \le p\}$. 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 $\{X_1 \ge x_1 \text{ or } \dots \text{ or } X_d \ge x_d\}$, 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,

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

where l is the *stable tail dependence function* (STDF) and the p_j's are the marginal exceedance probabilities, $p_j = \mathbb{P}(X_j \ge x_j)$. 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); De Haan & Ferreira (2007) 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 population 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 7.

Theorem 1.2. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ 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 $1 - \delta$,*

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

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

$$\sup_{0 \leq \mathbf{x} \leq T} |l_n(\mathbf{x}) - l(\mathbf{x})| \;\leq\; Cd\sqrt{\frac{T}{k} \log \frac{d+3}{\delta}} \;+\; \textit{bias}(k, n, T),$$

where l is the STDF*and l_n its standard empirical version. The second term in the bound is a bias stemming from the asymptotic nature of l.*

In this section, we have introduced and studied, in a non-parametric setting, a particular functional characterizing the extreme dependence structure. One other convenient (non-parametric) 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); 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.

1.4.2 Sparse Representation of Multivariate Extremes

This section is a summary of Chapter 8, which is based on previous work published in Goix et al. (2016c) and on its long version Goix et al. (2016b) under review.

EVT has been intensively used in anomaly detection in the one-dimensional situation, see for instance Roberts (1999, 2000); Clifton et al. (2011, 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.

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 (2014) 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 concomittantly 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); Einmahl & Segers (2009), in an asymptotic framework. Here we extend the non-asymptotic 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 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 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.

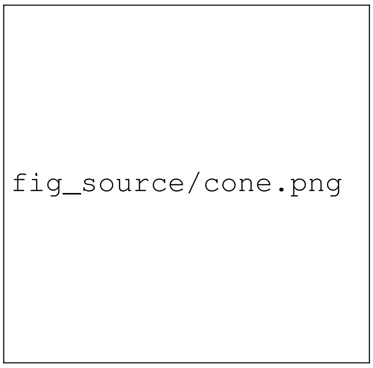


FIGURE 1.2: Truncated cones in 3D

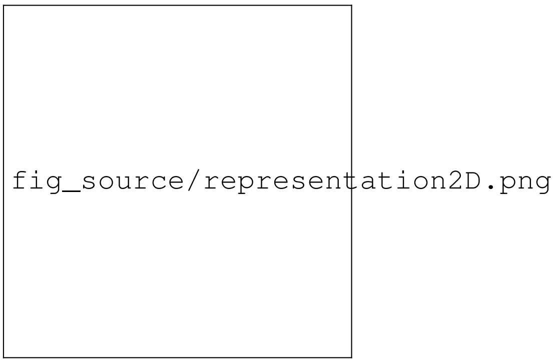


FIGURE 1.3: Truncated ϵ-cones in 2D

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

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

Theoretically, recovering the $(2^d - 1)$-dimensional unknown vector $\mathcal{M}$ amounts to roughly approximating the support of Φ using the partition $\{\Omega_\alpha, \alpha \subset \{1, \ldots, d\}, \alpha \neq \emptyset\}$, 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 8 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 subsection 1.4.1, in particular from Theorem 1.2 and from the ideas used to prove Theorem 1.3, 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

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

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

1.5 Heuristic approaches

The two contributions in this section are of heuristic nature and not yet supported by statistically sound theoretical results. Although this ongoing work has not been published yet and will certainly be completed in the near future, we believe that it has its place in our manuscript, given the numerous convincing numerical experiments we carried out and the rationale behind the approaches promoted we gave. These two contributions address two major challenges in anomaly detection:

- How to evaluate unsupervised anomaly detection in practice?
- How to grow random forests with only one class available?

The first point has been partially addressed in Section 1.3 with MV and EM curves. However, 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. Besides, their use 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 very challenging in high dimensional frameworks.

The two proposed approaches are heuristic-based, and no theoretical guarantees such as consistency or convergence rates are derived. However, extensive benchmarks show the relevance of these approaches.

1.5.1 Evaluation of anomaly detection algorithms

This is a summary of Chapter 9, which is based on a workshop paper (Goix, 2016) and a work to be submitted (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; Clémençon & Vayatis, 2009a) 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

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

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

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 $T : \text{Im(f)} \to \mathbb{R}$. As curves cannot be trivially compared, consider the L^1-norm $\|.\|_{L^1(I)}$ 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 $\mathbf{X}_1, \ \ldots, \ \mathbf{X}_n$ 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),

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

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

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

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

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

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 $\widehat{\mathcal{C}}_k^{MV}$ (resp. $\widehat{\mathcal{C}}_k^{EM}$) 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 $X = (x_i^j)_{1 \le i \le n, 1 \le j \le d}$, 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 $\widehat{s}_k = \mathcal{A}\big((x_i^j)_{1 \le i \le n,\ j \in F_k}\big)$
    compute $\widehat{\mathcal{C}}_k^{EM} = \|\widehat{EM}_{\widehat{s}_k}\|_{L^1(I^{EM})}$ using (1.10) or $\widehat{\mathcal{C}}_k^{MV} = \|\widehat{MV}_{\widehat{s}_k}\|_{L^1(I^{MV})}$ using (1.11)
**end for**
**Return** performance criteria:

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

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.

1.5.2 One-Class Random Forests

This is a summary of Chapter 10, which is based on work (Goix et al., 2016a) to be submitted.

Building accurate scoring functions by optimizing EM or MV criteria is very challenging in practice, just as building classifiers by optimizing the ROC curve (Clémençon & 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 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 $X : \Omega \to \mathbb{R}^d$ *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:

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

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 $\mathbf{N} \sim \text{Bin}\big(n, (1 - \alpha)\big)$, 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 $\mathcal{X}$. This assumption *will be removed* in the adaptive approach, where no prior distribution is assumed for the outliers.

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 One-Class-Model(n, α) defined above.

The idea is to make $\alpha(t) \to 1$, $n(t) \to \infty$ when the volume of node t goes to zero. In other words, instead of considering one fixed general model One-Class-Model(n, α), we adapt it to each node t, considering 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 One-Class-Model($n(t)$, $\alpha(t)$) allows to maintain a non-negligible expected proportion of outliers in the node to be split, even when its volume becomes close to zero. 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 $\big(n(t), \alpha(t)\big)$ is

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

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 $\mathcal{P}\big((1-\alpha)n\big)$, and that an infinite number of outliers have been hidden. In the two class framework, this corresponds to observing an infinite number of outliers distributed closely around, outside and inside the support of the normal distribution, breaking the curse of dimensionality when using uniformly distributed outliers (see Chapter 10 for details).

Remark 1.4 (**Basic idea behind the adaptive approach**). This work corresponds in fact to the following simple idea that allows us to split a node without examples of the second class. Each time we are looking for the best split for a node t, we simply replace (in the 2-class impurity decrease to be maximized) the second class proportion in the left node t_L by the proportion expectation $volume(t_L)/volume(t)$ (idem for the right node). It ensures that one child node tries to capture the maximum number of observations with a minimal volume, while the other child looks for the opposite.

Remark 1.5 (**No sampling**). The corresponding sampling method is the following: for each note t to be splitted containing n_t observations (inliers), generate n_t uniform outliers over the corresponding cell to optimize a two-class splitting criterion. We precisely *avoid sampling* the outliers by using the proportion expectation $volume(t_L)/volume(t)$.

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 population 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 10 formally defines the one-class splitting criteria and provides an extensive benchmark of state-of-the-art anomaly detection algorithms.

1.6 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 5.2.

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.

1.7 Conclusion and Scientific Output

The contributions of this thesis can be summarized as follows.

First, an adequate performance criterion called Excess-Mass curve is proposed (Section 1.3.3), 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. (AISTATS 2015).
  Authors: Goix, Sabourin, and Clémençon.

As a second 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 1.4.1). 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 Clémençon.

The third 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 1.4.2). This contribution also includes a multivariate EVT-based algorithm which returns a scoring functions defined in extreme regions. This directly applies to anomaly detection as an abnormality score. The corresponding publications are Goix et al. (2016c), Goix et al. (2015a) and Goix et al. (2016b):

- Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. (AISTATS 2016 and NIPS 2015 Workshop on Nonparametric Methods for Large Scale Representation Learning).
  Authors: Goix, Sabourin, and Clémençon.

[1] https://github.com/scikit-learn

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

As a fourth 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? (to be submitted).
  Authors: Goix and Thomas.

The fifth contribution of this thesis is to develop an efficient heuristic for building accurate scoring functions. This is done by generalizing random forests to one-class classification. The corresponding work (to be submitted) is Goix et al. (2016a):

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

As a last contribution, two classical anomaly detection algorithms have been implemented and merged on scikit-learn. They are used in this dissertation for empirical comparison purpose to attest the relevance of the forementionned 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 3 presents general results on measure concentration inequalities;

Chapter 4 provides a concise background on extreme value theory;

Chapter 5 reviews classical anomaly detection algorithms used in the benchmarks and provides illustrative examples with the scikit-learn library. It also presents relating code contributions.

- Part II deals with theoretical performance criteria for the anomaly ranking task:

  Chapter 6 presents the details on anomaly ranking and excess-mass curve, as summarized above Section 1.3;

- Part III focuses on EVT-based methods for anomaly detection:

  Chapter 7 deals with the stable tail dependence function as summarized above in Section 1.4.1;

  Chapter 8 describes how scoring functions can be build using EVT, as previously summarized in Section 1.4.2.

- Part IV gathers two efficient heuristic-based methodologies:

  Chapter 9 deals with the evaluation of anomaly detection algorithms, as summarized above Section 1.5.1;

  Chapter 10 presents the details (summarized above Section 1.5.2) on one-class random forests.

CHAPTER 2
Résumé des contributions en Français

2.1 Introduction

Une *anomalie*, du grec $alpha\ nu\ omega\ mu\ alpha\ lambda\ iota\ alpha$, aspérité, irrégularité, non semblable (an-homalos), désigne un écart par rapport à une certaine norme, reflétant le comportement attendu. On appelle *anomalie* l'objet qui induit cet espace, l'observation qui s'écarte de la normalité. Dans beaucoup de domaines, la situation suivante se présente: un expert cherche à prédire un phénomène sur la base d'observations antérieures. Le cas le plus fondamental est lorsque l'on veut prédire certaines caractéristiques binaires d'observations nouvelles, compte tenu des précédentes. Par exemple, on peut penser à un médecin voulant prédire si un nouveau patient présente ou non une certaine pathologie, en utilisant les données des patients précédents (comme l'âge, l'histoire, le sexe, la pression artérielle) associée à leur véritable **étiquette/label**: avoir ou non la pathologie en question). Ce cas est un exemple de *classification binaire*, où le médecin cherche à trouver une règle pour **prédire** l'étiquette d'un nouveau patient (ce dernier étant caractérisé par son dossier médical, contenant toutes les mesures qui lui ont été faites). Cette règle est appelée un *classifieur* et doit être construite, *apprise*, sur des dossiers médicaux précédents. Intuitivement, le classificateur prédit le même diagnostic pour des dossiers médicaux similaires, dans un sens qui doit être appris avec précision.

On peut distinguer deux cas. Si les étiquettes des patients antérieurs sont connues (porteur ou non de la pathologie), on dit que la tâche de classification est **supervisée**. Si les étiquettes de ces données d'entrainement sont inconnues, la classification est dite **non-supervisée**. Suivant notre exemple, le médecin doit trouver deux formes (ou cluster) distincts dans les données, correspondant aux deux étiquettes, "en bonne santé" - "malade", formes qui contiennent chacune des dossiers de patients similaires.

La détection d'anomalies survient lorsqu'une étiquette est fortement sous-représentée dans les données d'entrainement, par exemple si très peu de patients ont la pathologie dans les données d'entrainement. Ainsi, la **détection d'anomalies supervisée** se résume à la classification supervisée de classes fortement déséquilibrées. En ce qui concerne la **détection d'anomalie non supervisée** (également appelée simplement détection d'outliers), elle suppose généralement que la base de données a un modèle "normal" caché, et les anomalies sont des observations qui s'écartent de ce modèle. Le médecin veut trouver des dossiers médicaux qui s'écartent de la grande majorité de ceux de ses patients précédents. Sa tâche est en quelque sorte simplifiée s'il sait que tous ses patients antérieurs sont en bonne santé: il est plus facile pour lui d'apprendre le modèle "normal", c'est-à-dire le dossier médical typique d'un patient en bonne santé, à confronté avec les dossiers médicaux de ses nouveaux patients. Ce cadre est celui de la **détection de nouveauté** – également appelé classification à une classe ou détection d'anomalies semi-supervisées: les données d'entraînement ne contiennent que des instances normales.

Ce chapitre est organisé de la façon suivante. Section 2.2, la détection d'anomalie est formellement introduite, ainsi que la notion de fonction de score. Deux critères sur la qualité d'une fonction de score sont ensuites présentés section 2.3. La section 2.4 se concentre sur la théorie des valeurs extrêmes (EVT) pour gagner en précision sur les régions extrêmes. Après avoir introduit la STDF (stable tail deviation function) representant la structure de dépendance des événements rares (Section 2.4.1), on montre que la théorie des extrêmes multivariées peut être utile pour produire des fonctions de score précise sur les régions de faible probabilité (Section 2.4.2). La section 2.5 regroupe deux contributions de nature heuristique portant d'une part sur l'évaluation / la sélection d'algorithmes de détection d'anomalies non supervisés (Section 2.5.1) et d'autre part sur l'extension des forêts aléatoires à la classification à une classe (Section 2.5.2). La section 1.6 présente les contributions relatives à la librairie open-source scikit-learn. La section 1.7 énumère les productions scientifiques et conclut.

Notations A travers ce document, $\mathbb{N}$ désigne l'ensemble des entiers naturels, $\mathbb{R}$ and $\mathbb{R}_+$ désigne respectivement les ensembles des nombres réels et celui des nombres réels positifs. Les ensembles sont généralement écrit en lettre calligraphiques comme $\mathcal{G}$, et $|\mathcal{G}|$ désigne le nombre d'éléments dans $\mathcal{G}$.

Ls vecteurs sont écrits en minuscules et en gras. Pour un vecteur $\mathbf{x} \in \mathbb{R}^d$ et $i \in \{1, \ldots, d\}$, x_i désigne la i^{me} composante de $\mathbf{x}$. Le produit scalaire entre deux vecteurs est noté $\langle \cdot, \cdot \rangle$. $\| \cdot \|$ désigne une norme arbitraire (sur des vecteurs ou sur des matrices) et $\| \cdot \|_p$ la norme L_p.

Au long de cette thèse, $\mathbb{P}[A]$ représente la probabilité de l'évènement $A \in \Omega$, l'espace de probabilité sous-jacent étant $(\Omega, \mathcal{F}, \mathbb{P})$. Nous utilisons la notation $\mathbb{E}[X]$ pour indiquer l'espérance de la variable aléatoire X. $X \overset{d}{=} Y$ signifie que X et Y sont égales en distribution et $X_n \overset{d}{\to} Y$ signifie que (X_n) converge vers Y en distribution. Nous utilisons souvent l'abréviation $\mathbf{X}_{1:n}$ pour désigner un échantillon $i.i.d.$ $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$.

Les notations sont résumées Table 2.1.

2.2 Détection d'anomalies, ranking d'anomalies et fonctions de scores

D'un point de vue probabiliste, il existe différentes façons de modéliser les comportements normaux et anormaux, ce qui conduit à différentes méthodologies. Un modèle probabiliste naturel consiste à supposer deux processus de génération différents pour les données normales et anormales. Les données normales (resp. données anormales) sont générées selon une distribution F (respectivement G). La distribution sous-jacente générale est alors un mélange de F et G. L'objectif est de déterminer si une nouvelle observation $\mathbf{x}$ a été générée à partir de F ou de G. Le meilleur moyen de résoudre théoriquement ce problème est le test du rapport de vraisemblances, également appelé test de Neyman-Pearson. Si $(\mathrm{d}F/\mathrm{d}G)(\mathbf{x}) > t$ avec un certain seuil $t > 0$, alors $\mathbf{x}$ a été tiré de F. Sinon, $\mathbf{x}$ est tiré de G. Cela revient à estimer l' *ensemble de niveaux de densité* $\{\mathbf{x}, (\mathrm{d}F/\mathrm{d}G)(\mathbf{x}) > t\}$ (Schölkopf et al., 2001; Steinwart et al., 2005; Scott & Nowak, 2006; Vert & Vert, 2006). Comme les anomalies sont très rares, leur structure ne peut être observée dans les données, en particulier leur distribution G. Il est courant et commode de remplacer G dans le problème ci-dessus par la mesure de Lebesgue, de sorte qu'il se résume à l'estimation du niveau de densité de F. (Schölkopf et al., 2001; Scott & Nowak, 2006; Vert & Vert, 2006).

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
$\mathrm{Leb}(\cdot)$	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
$\lvert\mathcal{G}\rvert$	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 $a \in \mathbb{R}^d$
$\lfloor\cdot\rfloor$	integer part
$\langle\cdot,\cdot\rangle$	Inner product between vectors
$\lVert\cdot\rVert$	An arbitrary norm
$\lVert\cdot\rVert_p$	L_p norm
$A\Delta B$	symmetric difference between sets A and B
$(\Omega, \mathcal{F}, \mathbb{P})$	Underlying probability space
$\mathcal{S}$	functions $s : \mathbb{R}^d \to \mathbb{R}_+$ 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 $\mathcal{E}$
$Y_{(1)} \leq \ldots \leq Y_{(n)}$	order statistics of $Y_1, \dots, Y_n$
$\mathbf{X}_{1:n}$	An $i.i.d.$ sample $(\mathbf{X}_1, \dots, \mathbf{X}_n)$
$\mathbb{P}[\cdot]$	Probability of event
$\mathbb{E}[\cdot]$	Expectation of random variable
$\mathbf{Var}[\cdot]$	Variance of random variable

TABLE 2.1: Summary of notation.

Cela revient à supposer que les anomalies sont uniformément réparties sur le support de la distribution normale. Cette hypothèse est donc implicitement faite par une majorité d'ouvrages sur la détection de nouveauté. Nous observons les données dans $\mathbb{R}^d$ à partir de la classe normale seulement, avec une distribution sous-jacente F et avec une densité $f : \mathbb{R}^d \to \mathbb{R}$. Le but est d'identifier les caractéristiques de cette classe normale, telles que son support $\{\mathbf{x}, f(\mathbf{x}) > 0\}$ ou un certain niveau de densité fixé $\{\mathbf{x}, f(\mathbf{x}) > T\}$ avec $t > 0$ près de 0.

La détection d'anomalies non supervisée est souvent considérée comme un problème de classification à une classe, où les données d'entraînement sont polluées par quelques éléments de la classe anormale: elle fait appel à des algorithmes à une classe *robustes* aux anomalies.

Une idée naturelle pour estimer les ensembles de niveaux de densité est de calculer une estimation de la densité et de considérer les ensembles de niveaux associés (Tsybakov, 1997; Cuevas & Fraiman, 1997; Baillo et al., 2001; Baillo, 2003; Cadre, 2006; Rigollet & Vert, 2009; Mason & Polonik, 2009). La densité est généralement estimée à l'aide d'un estimateur à noyau non paramétrique ou d'un estimateur de maximum de vraisemblance à partir d'une famille paramétrique de fonctions. Mais ces méthodes ne s'adaptent pas bien à la grande dimension. D'une certaine manière, ces méthodes cherchent à capturer plus d'informations que nécessaire pour la tâche d'estimation d'ensemble de niveau, comme les propriétés locales de la densité qui sont inutiles pour cette tache. En effet, il s'avère que pour toute transformation croissante T, les ensembles de niveaux de $T \circ f$ sont exactement ceux de f. Ainsi, il suffit d'estimer n'importe quel représentant de la classe des transformées croissantes de f, pour obtenir des estimés d'ensemble de niveaux. Intuitivement, il suffit d'estimer le pré-ordre (le *scoring*) induit par f sur $\mathbb{R}^d$. Définissons une *fonction de score* comme toute fonction mesurable $s :\ \mathbb{R}^d \to \mathbb{R}_+$ intégrable par rapport à la mesure de Lebesgue $\text{Leb}(.)$ et $\mathcal{S}$ l'espace De toutes les fonctions de score. Toute fonction de score définit un pré-ordre sur $\mathbb{R}^d$ et donc un classement sur un ensemble de nouvelles observations. Ce classement peut être interprété comme un degré d'anormalité, plus $s(x)$ est petit, plus x est normal. Notons que la plupart des algorithmes de détection d'anomalie renvoient plus qu'une étiquette binaire, normale / anormale. Ils renvoient une fonction de score, qui peut être convertie en prédiction binaire, généralement en imposant un seuil basé sur sa distribution statistique.

Supposons que nous voulons apprendre une fonction de score s dont le pré-ordre induit est "proche" de celui de f, ou de manière équivalente dont les ensembles de niveaux induits sont proches de ceux de f. Le problème est de transformer cette notion de proximité en critère $\mathcal{C}$, les fonctions de score optimales s^* étant alors définies comme celles qui optimisent $\mathcal{C}$. Dans le cadre de l'estimation de la densité, la différence uniforme $\|f - \hat{f}\|$
$infty$ est un critère commun pour évaluer la qualité de l'estimation. Nous aimerions un critère similaire, mais qui est invariant par transformé croissante de $\hat{f}$. En d'autres termes, le critère doit être défini de telle sorte que la collection d'ensemble de niveaux d'une fonction de score optimale $s^*(x)$ coïncide avec celle liée à f, et toute transformation croissante de la densité devrait être optimale au sens de $\mathcal{C}$. Plus formellement, nous allons considérer $\mathcal{C}^{\Phi}(s) = \|\Phi(f)\|$ (au lieu de $\|s - f\|$) avec $\Phi : \mathbb{R} \to \mathbb{R}_+$ vérifiant $\Phi(T \circ s) = \Phi(s)$ pour toute fonction de score s et transformation croissante T. Ici $\Phi(s)$ désigne soit la courbe masse-volume MV_s de s, soit sa courbe en excès-masse EM_s, définies dans la section suivante.

Ce critère qui mesure la qualité d'une fonction de score est alors un outil pour construire / apprendre une bonne fonction de score. Selon le paradigme de la minimisation du risque empirique, une fonction de score est construite en optimisant une version empirique $\mathcal{C}_n(s)$ du critère sur un ensemble adéquat de fonctions de score $\mathcal{S}_0$ de complexité contrôlée (par exemple une classe de dimension VC finie).

La section suivante décrit deux critères fonctionnels au vue de la nature globale du problème, tout comme les courbes ROC(*Receiver Operating Characteristic*) et PR (*Précision-Rappel*, et qui sont admissibles par rapport aux exigences énumérées ci-dessus. Ces critères fonctionnels étendent en quelque sorte le concept de la courbe ROC au cadre non-supervisé.

Remarque 1. **Terminologie: détection d'anomalies, ranking d'anomalies.**
À proprement parler, les critères que nous recherchons sont des critères de *ranking* d'anomalies, de la même manière que la courbe ROC est essentiellement un critère de *ranking* bipartite.

En pratique comme mentionné ci-dessus, tous les algorithmes de détection d'anomalies sont candidats à la tâche de ranking d'anomalie. Ils produisent tous une fonction de score, même ceux qui traitent à l'origine du cadre de "classification des anomalies", c'est à dire cherchent à être optimal sur un seul ensemble de niveau ou pour un taux de faux positifs fixe.

Dans la littérature, la terminologie "détection d'anomalies" est largement utilisée, au lieu de la terminologie plus précise de "ranking d'anomalies". Par exemple, Liu et al. (2008) écrit "*Le but de la détection d'anomalie est de fournir un ranking qui reflète le degré d'anomalie*".

Dans le cadre de ce travail, nous optons de même pour la convention que la détection d'anomalies se réfère au ranking d'anomalies: si les labels sont disponibles pour l'étape d'évaluation, l'objectif est de maximiser l'aire sous la courbe ROC. Si aucune donnée labelisée n'est disponible, l'objectif est de maximiser les critères non-supervisées définis dans la section suivante.

2.3 M-estimation et critères de performance pour les fonctions de scores

Cette section est un résumé du chapitre **??**, qui est basé sur un travail publié en Goix et al. (2015c). Nous fournissons un bref aperçu du critère de la courbe masse-volume introduit dans Clémençon & Jakubowicz (2013), qui est basé sur la notion d'ensembles de volume minimum. Nous exposons ensuite les principaux inconvénients de cette approche et proposons un autre critère, la courbe d'excès de masse.

2.3.1 Ensembles à volume minimal

La notion d'ensemble à volume minimal (Polonik (1997); Einmahl & Mason (1992)) a été introduite pour décrire des régions où une variable aléatoire multivarié $\mathbf{X} \in \mathbb{R}^d$ se trouvent avec très grande ou très petite probabilité. Soit $\alpha \in (0,1)$, un ensemble à volume minimal Γ_α^* de masse au moins α est une solution du problème de minimisation sous contrainte

$$\min_{\Gamma \text{ borelien}} \text{Leb}(\Gamma) \text{ tel que } \mathbb{P}(\mathbf{X} \in \Gamma) \geq \alpha, \tag{2.1}$$

le minimum étant pris sur tous les sous-ensembles mesurables Γ de $\mathbb{R}^d$. On peut montrer que chaque niveau de densité est un ensemble à volume minimal pour une certaine masse et que la réciproque est vraie si la densité n'a pas de partie plate. Dans le reste de cette section, on suppose que F a une densité $f(x)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ satisfaisant les hypothèses suivantes:

$\mathbf{A_1}$ *La densité f est bornée.*

$\mathbf{A_2}$ *La densité f n'a pas de partie plate: $\forall c \geq 0$, $\mathbb{P}\{f(\mathbf{X}) = c\} = 0$.*

Sous les hypothèses précédentes, pour n'importe quel $\alpha \in (0,1)$, il existe un unique ensemble à volume minimal Γ_α^*, dont la masse est equale à α. La fonction quantile (généralisée) est alors définie par:

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

En outre, l'application λ^* est continue sur $(0,1)$ et uniformément continue sur $[0, 1-\epsilon]$ pour tout $\epsilon \in (0,1)$ - quand le support de F est compacte, la continuité uniforme est valable sur l'intervalle fermé $[0,1]$.

Les estimés $\widehat{\Gamma}_\alpha^*$ des ensembles à volume minimal sont construits en remplaçant la distribution de probabilité inconnue F par sa version empirique $F_n = (1/n)\sum_{i=1}^N \delta_{\mathbf{X}_i}$ et en restraignant l'optimisation à une collection $\mathcal{A}$ de sous-ensembles boréliens de $\mathbb{R}^d$. $\mathcal{A}$ est supposée être assez riche pour inclure tous les ensembles de niveaux de densité, ou au moins des approximations raisonnables de ceux-ci.

Dans Polonik (1997), les résultats limites sont prouvés pour le processus de quantile empirique généralisé $\{\text{Leb}(\widehat{\Gamma}^*_\alpha) - \lambda^*(\alpha)\}$ (sous l'hypothèse en particulier que $\mathcal{A}$ est une classe de Glivenko-Cantelli pour F). Dans Scott & Nowak (2006), il est proposé de remplacer le niveau α par $\alpha - \phi_n$ où ϕ_n joue le rôle d'un paramètre de tolérance (du même ordre que le supremum $\sup_{\Gamma \in \mathcal{A}} |F_n(\Gamma) - F(\Gamma)|$), la complexité de la classe $\mathcal{A}$ étant contrôlée par la dimension VC, afin d'établir des bornes. La version statistique du problème du volume minimal est alors

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

L'ensemble $\mathcal{A}$ de sous-ensembles boréliens de $\mathbb{R}^d$ offre dans l'idéal des avantages statistiques et computationnels, permettant une recherche rapide tout en étant suffisamment complexe pour capturer la géométrie des ensembles de niveaux de la densité - en d'autre termes, le "biais de modèle" $\inf_{\Gamma \in \mathcal{A}} \text{Leb}(\Gamma \Delta \Gamma^*_\alpha)$ doit être petit.

2.3.2 La courbe Masse-Volume

Soit $s \in \mathcal{S}$ une fonction de score. Comme défini en Clémençon & Jakubowicz (2013); Clémençon & Robbiano (2014), la courbe masse-volume de s est le tracé de la fonction

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

où H^{-1} désigne la pseudo-inverse de n'importe quelle cdf $H : \mathbb{R} \to (0,1)$ et où α_s et λ_s sont définis par

$$\begin{aligned} \alpha_s(t) &:= \mathbb{P}(s(\mathbf{X}) \geq t), \\ \lambda_s(t) &:= \text{Leb}(\{\mathbf{x} \in \mathbb{R}^d, s(\mathbf{x}) \geq t\}) \,. \end{aligned} \tag{2.2}$$

Ceci induit un ordre partiel sur l'ensemble de toutes les fonctions de score: s est préférée à s' si $MV_s(\alpha) \leq MV_{s'}(\alpha)$ pour tout $\alpha \in (0,1)$. De plus, la courbe masse-volume reste inchangée lors de l'application d'une transformation croissante sur s. On peut prouver que $MV^*(\alpha) \leq MV_s(\alpha)$ pour tout $\alpha \in (0,1)$ et toute fonction de score s, où $MV^*(\alpha)$ est la valeur optimale du problème de minimisation sous contrainte (2.1), à savoir

$$MV^*(\alpha) = \text{Leb}(\Gamma^*_\alpha) = \min_{\Gamma\ mes.} \text{Leb}(\Gamma) \text{ subject to } \mathbb{P}(\mathbf{X} \in \Gamma) \geq \alpha \,. \tag{2.3}$$

Sous les hypothèses $\mathbf{A_1}$ et $\mathbf{A_2}$, on peut montrer que la courbe MV^* est bien une courbe masse volume, associée à (toute transformation croissante de) la densité f à savoir: $MV^* = MV_f$.

L'objectif est alors de construire une fonction de score $\hat{s}$ en fonction des données d'entraînement $\mathbf{X}_1, ...\mathbf{X}_n$ telle que $MV_{\hat{s}}$ soit minimale partout, c'est-à-dire minimisant $\|MV_{\hat{s}} - MV^*\|_\infty := \sup_{\alpha \in [0,1]} |MV_{\hat{s}}(\alpha) - MV^*(\alpha)|$.

Pour ce faire, il faut d'abord estimer une collection d'ensembles à volume minimal relatifs aux masses cibles $0 < \alpha_1 < \ldots < \alpha_K < 1$ formant une subdivision de $(0,1)$ sur la base des données d'entrainement afin de définir $s = \sum_k \mathbb{1}_{\{x \in \hat{\Gamma}^*_{\alpha_k}\}}$. L'analyse se fait sous des hypothèses

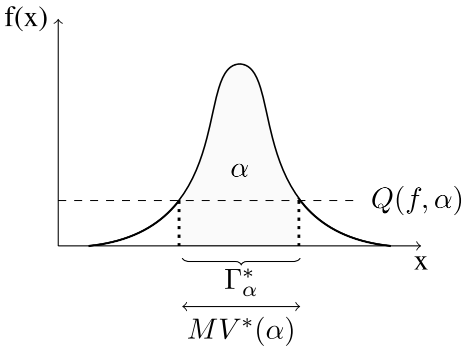


FIGURE 2.1: Masse-Volume au niveau α

adéquates (relatives à $\mathcal{G}$, au périmètre de $\Gamma^*_{\alpha_k}$ et au pas de la subdivision en particulier) et pour un choix approprié de $K = K_n$. Cependant, par construction, les vitesse d'apprentissage sont plutôt lentes (de l'ordre $n^{-1/4}$) et ne peuvent pas être établies lorsque le support n'est pas borné.

Les quatre principaux inconvénients de ce critère de courbe masse-volume sont les suivants.

1) Lorsqu'il est utilisé comme critère de performance, la mesure de Lebesgue d'ensembles pouvant être très complexes doit être calculée.

2) Lorsqu'il est utilisé comme critère de performance, il n'existe pas de méthode directe pour comparer les courbes MV puisque la zone sous la courbe est potentiellement infinie.

3) Lorsqu'il est utilisé comme critère d'apprentissage, il produit des ensembles de niveaux qui ne sont pas nécessairement imbriqués, puis des fonctions de notation imprécises.

4) Lorsqu'il est utilisé comme un critère d'apprentissage, les taux d'apprentissage sont plutôt lents (de l'ordre $n^{-1/4}$), et ne peuvent pas être établis dans le cas d'un support non borné.

Dans la section suivante, et comme contribution de cette thèse, un autre critère fonctionnel est proposé, obtenu en échangeant objectif et contrainte dans (2.1). Les inconvénients du critère de la courbe masse-volume sont résolus à l'exception du premier, et l'on montre que l'optimisation d'une version discrète empirique de cette mesure de performance donne des fonctions de score avec des taux de convergence de l'ordre $\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})$. En outre, les résultats peuvent être étendus à la situation où le support de la distribution F n'est pas compact. De plus, lorsqu'on relaxe l'hypothèse faite dans l'analyse de la courbe masse-volume que tous les vrais ensembles de niveaux sont inclus dans notre classe de minimisation $\mathcal{A}$, un contrôle du biais du modèle est établi. Enfin, nous déduisons des propriétés théoriques (non statistiques) vérifiées par ce critère, ce qui corrobore sa qualité de métrique sur les ensembles de niveaux contenus dans les fonctions de score.

2.3.3 Le critère d'excès de masse

Nous proposons un autre critère de performance qui s'appuie sur la notion de *d'excès de masse* et d' *ensemble de contours de densité*, comme introduits dans la contribution Polonik (1995).

L'idée principale est de considérer une formulation lagrangienne d'un problème de minimisation sous contrainte, obtenu en échangeant la contrainte et l'objectif dans (**??**): pour $t > 0$,

$$\max_{\Omega \text{ borelien}} \left\{ \mathbb{P}(\mathbf{X} \in \Omega) - t\text{Leb}(\Omega) \right\}. \tag{2.4}$$

On désigne par Ω_t^* une solution de ce problème. Cette formulation offre certains avantages à la fois computationnels et théoriques: en laissant (une version discrétisée) du multiplicateur lagrangien t augmenter de 0 à l'infini, on peut facilement obtenir des solutions à la contrepartie empirique de (**??**) formant une suite *imbriquée* d'ensembles, évitant ainsi une dégradation du taux de convergence due à la transformation des solutions empiriques pour forcer la monotonie.

La **courbe d'excès de masse optimale** d'une distribution de probabilité F est définie comme le graphe de la fonction

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

Avec les notations précédente, nous avons: $EM^*(t) = \mathbb{P}(\mathbf{X} \in \Omega_t^*) - t\text{Leb}(\Omega_t^*)$ pour tout $t > 0$. Remarquons que $EM^*(t) = 0$ pour tout $t > \|f\|_\infty := \sup_{x \in \mathbb{R}^d} |f(x)|$. La **courbe d'excès de masse** de $s \in \mathcal{S}$ par rapport à la distribution de probabilité F d'une variable aléatoire $\mathbf{X}$ est le graphe de la fonction

$$EM_s : t \in [0, \infty[\mapsto \sup_{A \in \{(\Omega_{s,l})_{l>0}\}} \{\mathbb{P}(\mathbf{X} \in A) - t\text{Leb}(A)\}, \tag{2.5}$$

où $\Omega_{s,t} = \{x \in \mathbb{R}^d, s(x) \geq t\}$ pour tout $t > 0$.

On peut également écrire EM_s en termes de λ_s et α_s définis en (2.2), $EM_s(t) = \sup_{u>0} \; \alpha_s(u) - t\lambda_s(u)$. Enfin, sous l'hypothèse $\mathbf{A_1}$, nous avons $EM_s(t) = 0$ pour tout $t > \|f\|_\infty$.

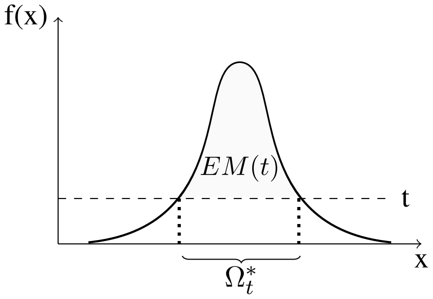


Figure 2: Excess-Mass curve

La maximisation de EM_s peut être vue comme trouver une collection de sous-ensembles $(\Omega_t^*)_{t>0}$ avec une masse maximale lorsqu'ils sont pénalisés par leur volume de façon linéaire. Une fonction de score optimale est alors n'importe quel $s \in \mathcal{S}$ admettant Ω_t^* 's comme ensembles de niveau, par exemple une fonction de score de la forme

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

avec $a(t) > 0$ (notons que $s(x) = f(x)$ pour $a \equiv 1$). La fonction EM_s est décroissante sur $(0, +\infty)$, à valeurs dans $[0, 1]$ et satisfait, $EM_s(t) \leq EM^*(t)$ pour tout $t \geq 0$. De plusn, pour

$t \geq 0$ et pour n'importe quel $\epsilon > 0$, nous avons

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

avec $\{s > u\} \Delta_\epsilon \{f > t\} \;:=\; \{f > t + \epsilon\} \setminus \{s > u\} \;\bigsqcup\; \{s > u\} \setminus \{f > t - \epsilon\}$. Ainsi la quantité $EM^*(t) - EM_s(t)$ mesure avec quelle qualité les ensembles de niveaux de s peuvent-il approchés ceux de la densité sous-jacente. Sous des hypothèse raisonnables, (voir Goix et al. (2015c), Prop.1), nous avons aussi pour $\epsilon > 0$,

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

où l'infimum est pris sur l'ensemble $\mathcal{T}$ de toutes les transformations croissantes mesurables $T : \mathbb{R}_+ \to \mathbb{R}_+$. Les inégalités précédentes révèlent que $\|EM^* - EM_s\|_\infty$ peut être interprété comme une pseudo distance, soit entre les ensembles de niveaux de s et ceux de la densité sous-jacente f, soit entre les pré-ordres induits par s et f.

Le concept de la courbe EM fournit un moyen simple de comparer les fonctions de score, mais l'optimisation d'un tel critère fonctionnel est loin d'être simple. Comme proposé dans Clémençon & Jakubowicz (2013) pour le critère MV, l'optimisation est faite sur une certaine classe de fonctions de score, que nous espérons assez riche pour fournir une bonne approximation (biais de modèle petit) tout en étant assez simple pour contrôler le taux de convergence. Nous considérons ici les fonctions de score de la forme

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

où $\mathcal{G}$ est une class VC de sous-ensembles de $\mathbb{R}^d$. Nous choisissons de manière arbitraire $a_k := (t_k - t_{k+1})$ de tel sorte que les $\hat{\Omega}_{t_k}$ correspondent exactement aux ensembles de niveau t_k, $\{s \geq t_k\}$. Ensuite, la maximisation du critère fonctionnel d'excès de masse s'effectue en résolvant de manière séquentielle, pour $k = 1, \ldots, N$,

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

Les solution $\hat{\Omega}_{t_k}$ de ces problèmes d'optimisation peuvent toujours être choisie de manière à être imbriquée (contrairement au problème d'optimisation analogue pour le critère masse-volume). En d'autres termes, une contrainte d'inclusion peut être incorporée dans le problème d'optimisation précédent, sans affecter la qualité de la solution obtenue.Dans le cadre du critère masse-volume, des hypothèses sont faites stipulant que le support de la distribution doit être compact et que la classe VC $\mathcal{G}$ doit contenir les vrais ensembles de niveaux de la densité. Ici, nous relaxons ces hypothèses, la première en choisissant des niveaux adaptatifs t_k, et la seconde en dérivant une étude de biais. Ceci est détaillé chapitre **??**.

2.4 Précision sur les régions extrêmes

2.4.1 Analyse du point de vue de la théorie des valeurs extrêmes par l'estimation de la STDF

Cette section est un résumé du chapitre 7, qui est basé sur le travail publié dans Goix et al. (2015b).

Rappelons que les fonctions de score sont construites en approchant les ensembles de niveaux de densité / ensembles à volume minimal de la densité "normale" sous-jacente. Comme nous l'avons mentionné précédemment, dans le cadre de la détection d'anomalie, nous souhaitons être précis sur des ensembles de niveaux correspondant à des quantiles élevés, à savoir avec un niveau t près de 0 – ou de manière équivalente, être précis sur des ensembles à volume minimal avec une contrainte de masse α proche de 1.

Dans le cas univarié, supposons que nous voulons considérer le quantile $(1-p)^{th}$ de la distribution F d'une variable aléatoire X, pour une probabilité donnée de dépassement p, c'est-à-dire $x_p = \inf\{x \in \mathbb{R},\ \mathbb{P}(X > x) \leq p\}$. Pour les valeurs modérées de p, une contrepartie empirique naturelle est $x_{p,n} = \inf\{x \in \mathbb{R},\ 1/n \sum_{i=1}^n \mathbb{1}_{X_i > x} \leq p\}$. Cependant, si p est très petit, l'échantillon fini $X_1, \ldots, X_n$ ne contient pas suffisamment d'informations $x_{p,n}$ devient inutile. Ce problème devient dans le cas multivarié celui d'estimer des ensembles de niveaux de densité avec un niveau très faible ou de manière equivalente celui d'estimer les fonctions de score associées à ces ensembles de niveaux. La théorie des valeurs extrêmes traite spécialement de ces problèmes, aussi bien dans le cadre unidimensionnel que multidimensionnel.

Preliminaries. La théorie de la valeur extrême (EVT) développe des modèles pour apprendre l'insolite plutôt que l'habituel. Ces modèles sont largement utilisés dans les domaines de la gestion des risques comme celui de la finance, de l'assurance, des télécommunications ou des sciences de l'environnement. Une application majeure de la EVT est de fournir une évaluation raisonnable de la probabilité d'occurrence d'événements rares.

Pour illustrer ce point, supposons que nous voulons gérer le risque d'un portefeuille contenant d actifs différents, $\mathbf{X} = (X_1, \ldots, X_d)$. Un but assez général est alors d'évaluer la probabilité d'événements du type $\{X_1 \geq x_1 \text{ or } \ldots \text{ or } X_d \geq x_d\}$, pour des seuils multivariés grands $\mathbf{x} = (x_1, \ldots, x_d)$. Dans des conditions pas trop strictes sur la régularité de la distribution $\mathbf{X}$, la EVT montre que pour des seuils suffisamment importants,

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

où l est la STDF *stable tail dependence function* et où les p_j sont les probabilités de dépassement marginal, $p_j = \mathbb{P}(X_j \geq x_j)$. La fonction l caractérise la *dépendance* entre les extrêmes. La distribution *jointe* (sur des seuils importants) peut donc être récupérée à partir de la connaissance des distributions marginales avec la STDF l. Dans la pratique, l peut être tirée de données "modérément extrêmes", typiquement les k 'plus grandes 'parmi un échantillon de taille n, avec $k \ll n$.

L'estimation des p_j peut s'effectuer suivant un chemin bien pavé: dans le cas univarié, la EVT consiste essentiellement à modéliser la distribution des maxima (*resp.* la queue supérieure) en tant qu'élément des familles paramétriques de Gumbel, Fréchet ou Weibull (*resp.* par une distribution de Pareto généralisée).

Par contre, dans le cas multivarié, il n'y a pas de paramétrisation fini-dimensionnelle de la structure de dépendance. Ce dernier étant caractérisé par le STDF, l'estimation de ce fonctionnel est l'un des principaux problèmes de la EVT multivariée. Les propriétés asymptotiques de la contrepartie empirique de la STDF ont été largement étudiées, voir Huang (1992); Drees & Huang (1998); Embrechts et al. (2000); De Haan & Ferreira (2007) pour le cas bivarié et Qi (1997); Einmahl et al. (2012) pour le cas général multivarié sous hypothèses de fluidité.

Cependant, aucune borne n'existe sur l'estimation non-asymptotique. La contribution résumée dans la section suivante et publiée dans Goix et al. (2015b) dérive de telles bornes non asymptotiques. Nos résultats ne nécessitent aucune hypothèse autre que l'existence de la STDF.

Apprentissage de la structure de dépendance des événements rares. Les inégalités de VC classiques visent à bourner la déviation des quantités empiriques par rapport aux quantités théoriques sur des classes d'ensemble relativement simples, appelées classes VC. Assez souvent, ces classes recouvrent tout le support de la distribution sous-jacente. Cependant, lorsqu'il s'agit d'événements rares, il est intéressant d'avoir de telles bornes sur une classe d'ensembles qui ne couvre qu'une région de faible probabilité et contient donc (très) peu d'observations. Cela donne des bornes plus fines, puisque seules les différences entre de très petites quantités sont impliquées dans l'analyse. Le point de départ est l'inégalité VC énoncée ci-dessous et prouvée dans le chapitre 7.

Theorem 2.1. *Soit $\mathbf{X}_1, \dots, \mathbf{X}_n$ des réalisations $i.i.d.$ d'une variable aléatoire $\mathbf{X}$, $\mathcal{A}$ une classe VC de VC-dimension $V_{\mathcal{A}}$.*

Considerons la classe $\mathbb{A} = \cup_{A \in \mathcal{A}} A$, et posons $p = \mathbb{P}(\mathbf{X} \in \mathbb{A})$. Alors il existe une constante absolue C de sorte que pour tout $0 < \delta < 1$, avec probabilité au moins $1 - \delta$,

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

L'idée principale est la suivante. L'estimateur empirique de la STDF est basé sur la mesure empirique des régions "extrêmes", qui sont touchées seulement avec une faible probabilité. Il suffit donc de borner les déviations maximales sur ces régions à faible probabilité. La clé consiste à choisir une classe VC adaptative, qui ne couvre que ces régions là, et à dériver des inégalités de type VC qui intègrent p, la probabilité de toucher la classe. La borne obtenue sur l'erreur non asymptotique est alors:

Theorem 2.2. *Soit T un nombre positif tel que $T \geq \frac{7}{2}(\frac{\log d}{k} + 1)$, δ tel que $\delta \geq e^{-k}$ et soit $k = k(n)$ une suite d'entiers strictement positifs telle que $k \to \infty$ et $k = o(n)$ quand $n \to \infty$. Alors il existe une constante absolue C telle que pour chaque $n > 0$, avec probabilité au moins $1 - \ delta$:*

$$\sup_{0 \leq \mathbf{x} \leq T} |l_n(\mathbf{x}) - l(\mathbf{x})| \leq Cd \sqrt{\frac{T}{k} \log \frac{d+3}{\delta}} + \textit{bias}(k, n, T),$$

où l est la stdf et l_n sa version empirique standard. Le second terme dans la borne est un biais issu de la nature asymptotique de l.

Dans cette section, nous avons introduit et étudié, dans un cadre non-paramétrique, une fonctionnelle particulière caractérisant la structure de dépendance des extrêmes. Une autre caractérisation pratique (non paramétrique) de cette dépendance dans le cadre de la EVT multivariée est la *mesure angulaire* qui fournit des informations directes sur les "directions" probables des extrêmes, c'est-à-dire la contribution relative de chaque coordonnée dans les "grandes" observations.

Dans de nombreuses applications, il est plus commode de travailler avec la mesure angulaire elle-même. Cette dernière donne des informations plus directes sur la structure de dépendance et est capable de refléter des propriétés structurelles (par exemple la sparsité/parcimonie comme détaillé ci-dessous) qui n'apparaîtraient pas dans les copules ou dans la STDF, ces derniers étant des versions intégrées de la mesure angulaire. Cependant, la modélisation non paramétrique de la mesure angulaire est confrontée à des difficultés majeures, dues à sa structure potentiellement complexe, en particulier dans un cadre de grande dimension. D'autre part, d'un point de vue théorique, l'estimation non paramétrique de la mesure angulaire n'a été étudiée que dans le cas bidimensionnel et dans un cadre asymptotique Einmahl et al. (2001); Einmahl & Segers (2009). La section ci-dessous résume une nouvelle méthodologie visant à représenter parcimonieusement dans la structure de dépendance des extrêmes.

2.4.2 Sparse Representation of Multivariate Extremes

Cette section est un résumé du chapitre 8, qui est lui-même basé sur les travaux précédents publiés en Goix et al. (2016c), ainsi que sur sa version longue Goix et al. (2016b) en cours de révision.

La EVT a été intensivement utilisé en détection d'anomalies dans le cas unidimensionnelle, voir par exemple Roberts (1999, 2000); Clifton et al. (2011, 2008); Lee & Roberts (2008). Dans le cas multivariée, cependant, il n'existe – à notre connaissance – aucune méthode de détection d'anomalies reposant sur la EVT *multivariée*. Jusqu'à présent, le cas multidimensionnel n'a été abordé que par l'usage de statistiques basées sur la EVT univariée. La raison majeure est la difficulté du passage à l'échelle des modèles multivariés avec la dimension. Dans le présent travail, nous comblons l'écart entre la détection d'anomalies et la EVT multivariée en proposant une méthode qui est capable d'apprendre un "profil normal" parcimonieux des extrêmes multivariés et, en tant que telle, peut être mise en œuvre pour améliorer la précision de tout algorithme de détection d'anomalies.

Context: Extrèmes multivariés en grande dimension. L'estimation paramétrique ou semi-paramétrique de la structure des extrêmes multivariés est relativement bien documentée dans la littérature, voir par exemple Coles & Tawn (1991); Fougères et al. (2009); Cooley et al. (2010); Sabourin & Naveau (2014) et leurs références. Cependant, des hypothèses structurelles restrictives doivent être faites, stipulant par exemple que seuls quelques sous-groupes prédéfinis de composantes peuvent être extrêmes ensemble, ou au contraire, que toutes doivent l'être. En outre, leur utilisation pratique est limitée à des problèmes en dimension modérée (par exemple, $d\ le 10$), sinon des choix de modélisation simplifiés sont nécessaires, comme dans Stephenson (2009)). Enfin, l'évaluation de l'incertitude concernant la production de ces modèles est faite sous l'hypothèse que les données d'entraînement sont "asymptotiques", au sens où l'on suppose que, quand elles excèdent un grand seuil fixé, les données sont exactement réparties selon la distribution limite des extrêmes . En d'autres termes, l'erreur de modélisation est ignorée.

L'estimation non-paramétrique de la mesure angulaire n'a été traitée que dans le cas bidimensionnel, dans Einmahl et al. (2001); Einmahl & Segers (2009), et dans un cadre asymptotique. Nous allons étendre l'étude non-asymptotique sur l'estimation de la STDF (section précédente) à la mesure angulaire des extrêmes, restreinte à une classe bien choisie d'ensembles. L'objectif est d'apprendre une représentation de la mesure angulaire, assez simple pour contrôler la variance en grande dimension et suffisamment précise pour obtenir des informations sur les "directions probables" des extrêmes. Ceci donne une première estimation non paramétrique de la mesure angulaire en dimension quelconque, limitée à une classe de sous-cones, avec une borne non asymptotique sur l'erreur. Notons que ce procédé peut également être utilisé comme étape de prétraitement, dans un cadre de réduction de dimension, avant de procéder à une estimation paramétrique ou semi-paramétrique qui pourrait bénéficier des informations de structure émises lors de la première étape. De telles applications dépassent le cadre de cette thèse.

Le cadre que nous développons est non paramétrique et se trouve à l'intersection de l'estimation de support, de l'estimation de densité et de la réduction de dimension: il consiste à apprendre le support d'une distribution (à partir des données d'apprentissage), qui peut être décomposé en sous-cones, potentiellement de dimension faible et auxquels une certaine masse est assignée.

Ceci produit une fonction de score définie sur les régions extrêmes, qui peut ainsi être exploitée pour détecter les anomalies parmi les extrêmes. En raison de sa complexité modérée - d'ordre $dn \log n$ - cet algorithme convient au traitement de problèmes d'apprentissage à grande échelle, et les résultats expérimentaux révèlent une performance significativement accrue sur les régions extrêmes par rapport aux approches de détection d'anomalies standard.

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 2.2 and 2.3 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 2.2.

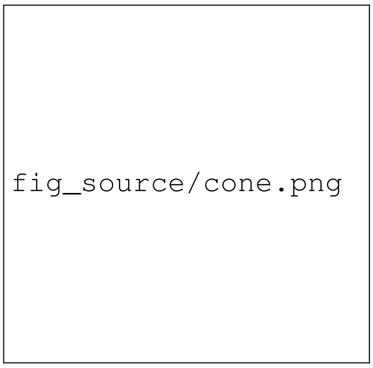


FIGURE 2.2: Truncated cones in 3D

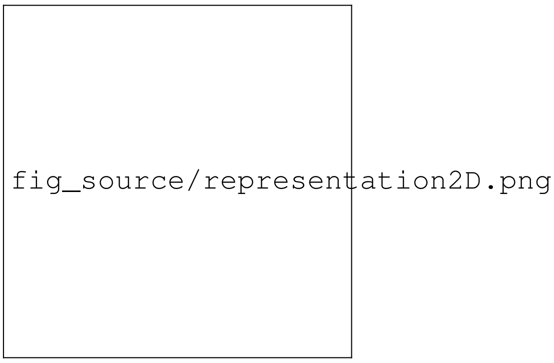


FIGURE 2.3: Truncated ϵ-cones in 2D

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 2.3 in the two-dimensional case. We thus obtain an estimate $\widehat{\mathcal{M}}$ of the representation

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

Theoretically, recovering the $(2^d - 1)$-dimensional unknown vector $\mathcal{M}$ amounts to roughly approximating the support of Φ using the partition $\{\Omega_\alpha, \alpha \subset \{1, \ldots, d\}, \alpha \neq \emptyset\}$, 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 8 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 subsection 2.4.1, in particular from Theorem 2.1 and from the ideas used to prove Theorem 2.2, 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

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

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

2.5 Heuristic approaches

The two contributions in this section are of heuristic nature and not yet supported by statistically sound theoretical results. Although this ongoing work has not been published yet and will certainly be completed in the near future, we believe that it has its place in our manuscript, given the numerous convincing numerical experiments we carried out and the rationale behind the approaches promoted we gave. These two contributions address two major challenges in anomaly detection:

- How to evaluate unsupervised anomaly detection in practice?
- How to grow random forests with only one class available?

The first point has been partially addressed in Section 2.3 with MV and EM curves. However, 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. Besides, their use 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 very challenging in high dimensional frameworks.

The two proposed approaches are heuristic-based, and no theoretical guarantees such as consistency or convergence rates are derived. However, extensive benchmarks show the relevance of these approaches.

2.5.1 Evaluation of anomaly detection algorithms

This is a summary of Chapter 9, which is based on a workshop paper (Goix, 2016) and a work to be submitted (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; Clémençon & Vayatis, 2009a) 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

$$MV_s(\alpha) = \inf_{u \geq 0} \ \text{Leb}(s \geq u) \ \textit{s.t.} \ \mathbb{P}(s(\mathbf{X}) \geq u) \geq \alpha \tag{2.6}$$

$$EM_s(t) = \sup_{u \geq 0} \ \mathbb{P}(s(\mathbf{X}) \geq u) \ - \ t\text{Leb}(s \geq u) \tag{2.7}$$

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 $T : \text{Im(f)} \to \mathbb{R}$. As curves cannot be trivially compared, consider the L^1-norm $\|.\|_{L^1(I)}$ 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 $\mathbf{X}_1,\ \ldots,\ \mathbf{X}_n$ 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 (2.6) and (2.7),

$$\widehat{MV}_s(\alpha) = \inf_{u \geq 0} \{\text{Leb}(s \geq u) \ \textit{s.t.}\ \mathbb{P}_n(s \geq u) \geq \alpha\} \tag{2.8}$$

$$\widehat{EM}_s(t) = \sup_{u \geq 0} \mathbb{P}_n(s \geq u) \ - \ t\text{Leb}(s \geq u) \tag{2.9}$$

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

$$\widehat{\mathcal{C}}^{EM}(s) = \|\widehat{EM}_s\|_{L^1(I^{EM})} \qquad I^{EM} = [0, \widehat{EM}^{-1}(0.9)], \tag{2.10}$$

$$\widehat{\mathcal{C}}^{MV}(s) = \|\widehat{MV}_s\|_{L^1(I^{MV})} \qquad I^{MV} = [0.9, 0.999]. \tag{2.11}$$

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 $F_1, \ldots, F_m$. A partial score $\widehat{\mathcal{C}}_k^{MV}$ (resp. $\widehat{\mathcal{C}}_k^{EM}$) is computed for each draw F_k using (2.10) (resp. (2.11)). The final performance criteria are obtained by averaging these partial criteria along the different draws of features. This methodology is described in Algorithm 2.

Algorithm 2 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 $k = 1, \ldots, m$ do
    randomly select a sub-group $F_k$ of $d'$ features
    compute the associated scoring function $\widehat{s}_k = \mathcal{A}\big((x_i^j)_{1 \leq i \leq n,\ j \in F_k}\big)$
    compute $\widehat{\mathcal{C}}_k^{EM} = \|\widehat{EM}_{\widehat{s}_k}\|_{L^1(I^{EM})}$ using (2.10) or $\widehat{\mathcal{C}}_k^{MV} = \|\widehat{MV}_{\widehat{s}_k}\|_{L^1(I^{MV})}$ using (2.11)
end for
Return performance criteria:

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

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.

2.5.2 One-Class Random Forests

This is a summary of Chapter 10, which is based on work (Goix et al., 2016a) to be submitted.

Building accurate scoring functions by optimizing EM or MV criteria is very challenging in practice, just as building classifiers by optimizing the ROC curve (Clémençon & 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 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 $X : \Omega \to \mathbb{R}^d$ *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:

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

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 $\mathbf{N} \sim \text{Bin}\big(n, (1 - \alpha)\big)$, 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 $\mathcal{X}$. This assumption *will be removed* in the adaptive approach, where no prior distribution is assumed for the outliers.

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 One-Class-Model(n, α) defined above.

The idea is to make $\alpha(t) \to 1$, $n(t) \to \infty$ when the volume of node t goes to zero. In other words, instead of considering one fixed general model One-Class-Model(n, α), we adapt it to each node t, considering 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 One-Class-Model($n(t)$, $\alpha(t)$) allows to maintain a non-negligible expected proportion of outliers in the node to be split, even when its volume becomes close to zero. 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 $\big(n(t), \alpha(t)\big)$ is

$$(1-\alpha)n = \big(1-\alpha(t)\big) \cdot n(t) \quad \text{for all } \ t, \tag{2.12}$$

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 $\mathcal{P}\big((1-\alpha)n\big)$, and that an infinite number of outliers have been hidden. In the two class framework, this corresponds to observing an infinite number of outliers distributed closely around, outside and inside the support of the normal distribution, breaking the curse of dimensionality when using uniformly distributed outliers (see Chapter 10 for details).

Remarque 2 (**Basic idea behind the adaptive approach**). This work corresponds in fact to the following simple idea that allows us to split a node without examples of the second class. Each time we are looking for the best split for a node t, we simply replace (in the 2-class impurity decrease to be maximized) the second class proportion in the left node t_L by the proportion expectation $volume(t_L)/volume(t)$ (idem for the right node). It ensures that one child node tries to capture the maximum number of observations with a minimal volume, while the other child looks for the opposite.

Remarque 3 (**No sampling**). The corresponding sampling method is the following: for each note t to be splitted containing n_t observations (inliers), generate n_t uniform outliers over the corresponding cell to optimize a two-class splitting criterion. We precisely *avoid sampling* the outliers by using the proportion expectation $volume(t_L)/volume(t)$.

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 population 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 10 formally defines the one-class splitting criteria and provides an extensive benchmark of state-of-the-art anomaly detection algorithms.

2.6 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 5.2.

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.

2.7 Conclusion and Scientific Output

The contributions of this thesis can be summarized as follows.

[1] https://github.com/scikit-learn

First, an adequate performance criterion called Excess-Mass curve is proposed (Section 2.3.3), 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. (AISTATS 2015).
  Authors: Goix, Sabourin, and Clémençon.

As a second 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 2.4.1). 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 Clémençon.

The third 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 2.4.2). This contribution also includes a multivariate EVT-based algorithm which returns a scoring functions defined in extreme regions. This directly applies to anomaly detection as an abnormality score. The corresponding publications are Goix et al. (2016c), Goix et al. (2015a) and Goix et al. (2016b):

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

As a fourth 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? (to be submitted).
  Authors: Goix and Thomas.

The fifth contribution of this thesis is to develop an efficient heuristic for building accurate scoring functions. This is done by generalizing random forests to one-class classification. The corresponding work (to be submitted) is Goix et al. (2016a):

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

As a last contribution, two classical anomaly detection algorithms have been implemented and merged on scikit-learn. They are used in this dissertation for empirical comparison purpose to attest the relevance of the forementionned 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 3 presents general results on measure concentration inequalities;

  Chapter 4 provides a concise background on extreme value theory;

  Chapter 5 reviews classical anomaly detection algorithms used in the benchmarks and provides illustrative examples with the scikit-learn library. It also presents relating code contributions.

- Part II deals with theoretical performance criteria for the anomaly ranking task:

  Chapter 6 presents the details on anomaly ranking and excess-mass curve, as summarized above Section 2.3;

- Part III focuses on EVT-based methods for anomaly detection:

  Chapter 7 deals with the stable tail dependence function as summarized above in Section 2.4.1;

  Chapter 8 describes how scoring functions can be build using EVT, as previously summarized in Section 2.4.2.

- Part IV gathers two efficient heuristic-based methodologies:

  Chapter 9 deals with the evaluation of anomaly detection algorithms, as summarized above Section 2.5.1;

  Chapter 10 presents the details (summarized above Section 2.5.2) on one-class random forests.

PART I

Preliminaries

CHAPTER 3
Concentration Inequalities from the Method of bounded differences

Abstract This chapter presents general results on measure concentration inequalities, obtained via martingale methods or with Vapnik-Chervonenkis theory. In the last section 3.4 of this chapter, a link is also made with contributions presented in Chapter 7 which builds on some concentration inequality stated and proved here.

Note: In addition to a review of popular results and sketches of proofs from the existing literature, the last section 3.4 of this chapter presents an original contribution; a VC-type inequality is proved using a Bernstein-type concentration inequality. A corollary of this VC-type inequality focusing on maximal deviation on low-probability regions is needed in Chapter 7.

We recommend McDiarmid (1998) and Janson (2002) for good references on this subject, and Massart (2007); Boucheron et al. (2013) for a extensive review on concentration inequalities. About the impact of the concentration of measure phenomenon in EVT, see Boucheron & Thomas (2012, 2015) and the PhD thesis Thomas (2015). References on classification and statistical learning theory (not gathered by this background part), include Vapnik & Chervonenkis (1974); Devroye et al. (1996); Bousquet et al. (2004); Boucheron et al. (2005); Bishop (2006); Friedman et al. (2001); Vapnik (2013)

3.1 Two fundamental results

The two theorems 3.1 and 3.4 presented in this section are powerful and allow to derive many classical concentration inequalities, like Hoeffding, Azuma, Bernstein or McDiarmid ones. The first theorem 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 $\mathcal{F}$.

Notation 1. Let us assume that X is a real *r.v.* and that $X \in L^\infty(\Omega)$. The conditional essential supremum $\sup(X|\mathcal{G})$ is the (almost surely) unique real *r.v.* $f : \Omega \to \mathbb{R}$ satisfying:

(i) f is $\mathcal{G}$-measurable

(ii) $X \leq f$ a.s.

(iii) If $g : \Omega \to \mathbb{R}$ 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 $\mathcal{G}_1 \subset \mathcal{G}_2$. For more properties enjoyed by conditional essential suprema, see Barron et al. (2003).

Notation 2. We still assume that X is a bounded *r.v.*. Let $(\mathcal{F}_k)_{0\leq k\leq n}$ be a filtration of $\mathcal{F}$ such that X is $\mathcal{F}_n$-measurable. We denote $X_1, ..., X_n$ 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 denote:

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

Notation 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 denote:

- $\mathbf{var_k} = var(Y_k|\mathcal{F}_{k-1}) = var(X_k|\mathcal{F}_{k-1})$ the conditional variance,
- $\mathbf{V} = \sum_1^n var_k$ the sum of conditional variances and $\hat{\nu} = \operatorname{ess\,sup}(V)$ the maximum sum of conditional variances,
- ∗ $\mathbf{dev_k^+} = \sup(Y_k|\mathcal{F}_{k-1})$ the conditional positive deviation,
- ∗ $\mathbf{maxdev^+} = \operatorname{ess\,sup}(\max_{0\leq k\leq n} dev_k^+)$ 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 $\mathbb{E}(V) = var(X)$.

3.1.2 Inequality for Bounded Random Variables

Theorem 3.1. *(McDiarmid, 1998) Let X be 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 $\mathcal{F}_n$-measurable. Then for any $t \geq 0$,*

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

and more generally

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

To prove this result the following lemmas are needed.

Lemma 3.2. *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*

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

Proof. This result can be easily proved by induction.

$$\begin{aligned}
\mathbb{E}\left[Ze^{h\sum Y_k}\right] &= \mathbb{E}\left[e^{hY_1}\mathbb{E}\left[Ze^{h\sum_2^n Y_k} \mid \mathcal{F}_1\right]\right] \\
&= \mathbb{E}\left[e^{hY_1}\mathbb{E}\left[e^{hY_2}\ldots\, \mathbb{E}\left[\mathbb{E}\left[Z \mid \mathcal{F}_n\right] e^{hY_n} \mid \mathcal{F}_{n-1}\right] \ldots \mid \mathcal{F}_1\right]\right] \\
&\leq \mathbb{E}\left[e^{hY_1}\mathbb{E}\left[e^{hY_2}\ldots\, \mathbb{E}\left[\sup\left[Z \mid \mathcal{F}_n\right] e^{hY_n} \mid \mathcal{F}_{n-1}\right] \ldots \mid \mathcal{F}_1\right]\right] \\
&= \mathbb{E}\left[e^{hY_1}\mathbb{E}\left[e^{hY_2}\ldots\, \sup\left[Z\mathbb{E}\left[e^{hY_n} \mid \mathcal{F}_{n-1}\right] \mid \mathcal{F}_n\right] \ldots \mid \mathcal{F}_1\right]\right] \\
&= \sup\left[Z\prod_k \mathbb{E}(e^{hY_k}|\mathcal{F}_{k-1}) \mid \mathcal{F}_n\right] \\
&\leq \sup\left[Z\prod_k \mathbb{E}(e^{hY_k}|\mathcal{F}_{k-1})\right] \qquad \text{(since } \mathcal{F}_0 \subset \mathcal{F}_n\text{)}
\end{aligned}$$

□

This lemma allows to decompose the expectation of a product into (the supremum of) a product of expectations, although $\sum Y_k$ is not a sum of independent variables.

Lemma 3.3. *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 3.1. This proof follows a traditional scheme, based on four steps: Chernoff method (exponential Markov inequality introducing a parameter h); decomposition of the exponential term using independence (or in the present case using Lemma 3.2 which plays the same role); upper bound on each term with Lemma 3.3; and finally optimization in parameter h.

Let $X_k = \mathbb{E}(X|\mathcal{F}_{k-1})$ 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 $h > 0$,

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

From Lemma 3.3, $\mathbb{E}(e^{hY_k}|\mathcal{F}_{k-1}) \leq e^{\frac{1}{8}h^2 r_k^2}$ so that using Lemma 3.2,

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

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 3.4. *(McDiarmid, 1998) 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 $b = maxdev^+$ the maximum conditional deviation assumed to be finite, and $\hat{\nu} = \operatorname{ess\,sup} V$ the maximum sum of conditional variances also assumed to be finite. Then, for any $t \geq 0$,*

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

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

To prove this result, two lemmas are needed: Lemma 3.2 previously stated, exploiting the decomposition into martingale differences and thus playing the same role as independence; and the following lemma replacing Lemma 3.3 in the case of non-necessarily bounded *r.v.*, but with bounded variance.

Lemma 3.5. *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 3.4. The proof follows the same classical lines as the one of Theorem 3.1. Let $Y_1, ..., Y_n$ 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 3.5, $\mathbb{E}(e^{hY_k}|\mathcal{F}_{k-1}) \leq e^{h^2 g(hdev_k^+) var_k} \leq e^{h^2 g(hb) var_k}$ so that from Lemma 3.2 we obtain,

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

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 Popular Inequalities

In this section, we illustrate the strength of Theorem 3.1 and Theorem 3.4 by deriving as corollaries classical concentration inequalities. The first three propositions hold for bounded random variables and derive from Theorem 3.1. The last one (Bernstein) holds under variance assumption and derives from Theorem 3.4.

Proposition 3.6. *(*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 3.1 with $X = \sum_1^n Y_k$, $\mathcal{F}_k = \sigma(Y_1, ..., Y_k)$ 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, $ran_k = ran(Y_k|\mathcal{F}_k) \leq 2c_k$, hence $R^2 \leq 4\sum c_k^2$ and $\hat{r}^2 \leq 4\sum c_k^2$. By Theorem 3.1, $\mathbb{P}(X \geq t) \leq e^{\frac{-2t^2}{\hat{r}^2}} \leq e^{-\frac{t^2}{2\sum c_k^2}}$. Applying this inequality to $-X$, we obtain the desired result. □

Proposition 3.7. *(*MCDIARMID INEQUALITY, OR 'INDEPENDENT BOUNDED DIFFERENCES INEQUALITY'*) Let $X = (X_1, ..., X_n)$ where the X_i's are independent* r.v. *with respected values in A_i. Let $f : \prod A_k \to \mathbb{R}$ verifying the following Lipschitz condition.*

For any $x,\ x' \in \prod_1^n A_k,\quad |f(x) - f(x')| \leq c_k\quad$ *if* $\ x_j = x'_j,\ $ *for* $\ j \neq k,\ \ 1 \leq j \leq n.$ (3.1)

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

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

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 3.1 we have

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

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

Proposition 3.8. *(*HOEFFDING INEQUALITY*) Let $X_1, ..., X_n$ 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 3.7) with $A_k = [a_k, b_k]$, $f(x) = \sum x_k$ and $c_k = b_k - a_k$. Within this setting, $\hat{r}^2 \leq b_k - a_k$. □

Remark 3.9. This result can be directly proved with the classical lines as in Theorem 3.1: Exponential Markov inequality, sum of independent variables assumption (or of martingale differences), and use of Lemma 3.3 before optimization in h:

$$\begin{aligned}
\mathbb{P}(S_n - \mu \geq t) &\leq \mathbb{E}(e^{h(S_n - \mu)})e^{-ht} \\
\mathbb{E}(\prod e^{h(X_k - \mathbb{E}X_k)}) &= \prod \mathbb{E}(e^{h(X_k - \mathbb{E}X_k)}) \quad \text{(from independence)} \\
&\leq e^{\frac{1}{8}h^2 \sum(b_k - a_k)^2} \qquad \text{(from Lemma 3.3)},
\end{aligned}$$

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

Remark 3.10. Comparing the two previous McDiarmid and Hoeffding inequalities with Theorem 3.1, 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 allow to bound $\hat{r}^2$.

The three previous propositions ignore information about the variance of the underlying process. The following inequality deriving from Theorem 3.4 provides an improvement in this respect.

Proposition 3.11. *(*BERNSTEIN INEQUALITY*) Let $X_1, ..., X_n$ 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 $V = var(S_n)$ 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 3.12. (GAIN WITH RESPECT TO INEQUALITIES WITHOUT VARIANCE TERM) Assume that $0 \leq X_i \leq 1$ and consider renormalized quantities, namely $\tilde{S}_n := S_n/n$, $\tilde{\mu} := \mu/n$, $\tilde{V} = V/n^2$. Then,

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

with t typically of order between $1/n$ and $1/\sqrt{n}$. Thus, if the variance $\tilde{V}$ is small enough, Bernstein inequality 'almost' 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 used for deriving concentration bounds on low probability regions.

Proof. Let $F_k = \sigma(X_1, ..., X_n)$, $X = \sum(X_k - \mathbb{E}X_k) = S_n - \mu$, $\tilde{X}_k = \mathbb{E}(X|\mathcal{F}_k) = \sum_1^k (X_i - \mathbb{E}X_i)$ and $Y_k = \tilde{X}_k - \tilde{X}_{k-1}$. Then $Y_k = X_k - \mathbb{E}X_k$, hence $dev_k^+ \leq b$, $maxdev^+ \leq b$ and $var_k = var(Y_k|\mathcal{F}_{k-1}) = \mathbb{E}((Y_k - \mathbb{E}(Y_k|\mathcal{F}_{k-1}))^2|\mathcal{F}_{k-1}) = \mathbb{E}((Y_k - \mathbb{E}Y_k)^2) = var(Y_k)$. Therefore $\hat{\nu} = \operatorname{ess\,sup}(\sum var_k) = \operatorname{ess\,sup}(V) = V$. Theorem 3.4 applies and yields,

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

□

3.3 Connections with Statistical Learning and VC theory

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

$$f(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \sup_{A \in \mathcal{A}} \left| \mathbb{P}(\mathbf{X} \in A) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A} \right| , \tag{3.2}$$

where $\mathbf{X}_1, ..., \mathbf{X}_n$ 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*,

$$S_{\mathcal{A}}(n) = \max_{x_1, \ldots, x_d \in \mathbb{R}^d} |\{\{x_1, \ldots, x_n\} \cap A,\ A \in \mathcal{A}\}|, \tag{3.3}$$

can be bounded in that way,

$$S_{\mathcal{A}}(n) \leq (n+1)^{V_{\mathcal{A}}}, \tag{3.4}$$

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 3.7), with $c_k = 1/n$. Thus, Proposition 3.7 yields

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

or equivalently

$$f(\mathbf{X}_1, \dots, \mathbf{X}_n) \leq \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}} \tag{3.5}$$

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 $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = 0) = 1/2$. Then, the following result holds true.

Lemma 3.13. *Let* $\mathbf{X}_1, ..., \mathbf{X}_n$ $i.i.d.$ *random variables, and a VC-class* $\mathcal{A}$ *with VC-dimension* $V_{\mathcal{A}}$. *The following inequalities hold true:*

$$\begin{aligned} (i) \quad & \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n) \ \leq \ 2\mathcal{R}_n \\ (ii) \quad & \mathcal{R}_n \ \leq \ C\sqrt{\frac{V_{\mathcal{A}}}{n}} \end{aligned}$$

Remark 3.14. Note that bound (ii) holds even for the conditional Rademacher average

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

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 *symmetrization* 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:

$$
\begin{aligned}
&\mathbb{E} f(\mathbf{X}_1,\dots,\mathbf{X}_n)\\
&= \mathbb{E}\sup_{A\in\mathcal{A}}\left|\mathbb{P}(\mathbf{X}\in A) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{X}_i\in A}\right|\\
&= \mathbb{E}\sup_{A\in\mathcal{A}}\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{X}'_i\in A}\right] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{X}_i\in A}\right|\\
&= \mathbb{E}\sup_{A\in\mathcal{A}}\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{X}'_i\in A} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{X}_i\in A} \;\middle|\; \mathbf{X}_1,\dots,\mathbf{X}_n\right]\right|\\
&\le \mathbb{E}\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{X}'_i\in A} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{X}_i\in A}\right| \qquad (\text{since } \mathbb{E}\sup(.) \ge \sup\mathbb{E}(.))\\
&= \mathbb{E}\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\mathbb{1}_{\mathbf{X}'_i\in A} - \mathbb{1}_{\mathbf{X}_i\in A}\right)\right| \qquad (\text{since } \mathbb{1}_{\mathbf{X}'_i\in A} - \mathbb{1}_{\mathbf{X}_i\in A} \overset{\mathcal{L}}{=} \sigma_i(\mathbb{1}_{\mathbf{X}'_i\in A} - \mathbb{1}_{\mathbf{X}_i\in A}))\\
&\le \mathbb{E}\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathbb{1}_{\mathbf{X}'_i\in A}\right| \;+\; \sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^{n}-\sigma_i\mathbb{1}_{\mathbf{X}_i\in A}\right|\\
&= 2\mathcal{R}_n
\end{aligned}
$$

■

□

Thus, combining Lemma 3.13 with (3.5) we obtain the following version of the popular Vapnik-Chervonenkis inequality.

Theorem 3.15. (VAPNIK-CHERVONENKIS) *Let* $\mathbf{X}_1, ..., \mathbf{X}_n$ $i.i.d.$ *random variables, and a VC-class* $\mathcal{A}$ *with VC-dimension* $V_{\mathcal{A}}$. *Recall that* $f(\mathbf{X}_1,\dots,\mathbf{X}_n)$ *– see* (3.2) *– refers to the maximal deviation* $\sup_{A\in\mathcal{A}}\left|\mathbb{P}(\mathbf{X}\in A) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{X}_i\in A}\right|$. *For* $\delta > 0$, *with probability higher than* $1-\delta$*:*

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

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

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

The latter comes from the sub-optimal inequality $\mathcal{R}_n \le C\sqrt{\frac{V_{\mathcal{A}}\log(n)}{n}}$.

3.4 Sharper VC-bounds through a Bernstein-type inequality

In this section, we prove a refinement of Theorem 3.15 above, stated in Proposition 3.18. This result is useful for the study of maximal deviations on low-probability regions, see Chapter 7.

Contributions presented in Chapter 7 include some VC-type inequality obtained by using a Bernstein-type inequality instead of the McDiarmid one used in (3.5). As mentioned above, the variance of $f(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ 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 3.4. The following lemma guarantees that the latter applies.

Lemma 3.16. *Consider the* r.v. $f(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ *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

$$\begin{aligned} h(\mathbf{x}_1, \ldots, \mathbf{x}_k) = \mathbb{E}\left[f(\mathbf{X}_1, \ldots, \mathbf{X}_n)|\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_k = \mathbf{x}_k\right] \\ - \mathbb{E}\left[f(\mathbf{X}_1, \ldots, \mathbf{X}_n)|\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_{k-1} = \mathbf{x}_{k-1}\right] \end{aligned}$$

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

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

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

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

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

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

Considering the definition of f, we have:

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

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

$$\left|h(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \mathbf{x}_k)\right| \;\leq\; \mathbb{E} \sup_{A\in\mathcal{A}} \frac{1}{n} \left|\mathbb{1}_{\mathbf{x}_k\in A} - \mathbb{1}_{\mathbf{X}_k\in A}\right| \, . \tag{3.6}$$

The term on the right hand side of (3.6) is less than $\frac{1}{n}$ so that maxdev$^+ \leq \frac{1}{n}$. Moreover, if $\mathbf{X}'$ is an independent copy of $\mathbf{X}$, (3.6) yields

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

so that

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

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. □

Remark 3.17. (ON PARAMETER q) The quantity $q = \mathbb{E}\left(\sup_{A \in \mathcal{A}} |\mathbb{1}_{\mathbf{X}' \in A} - \mathbb{1}_{\mathbf{X} \in A}|\right)$ measures the complexity of class $\mathcal{A}$ **with respect to the distribution of 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* $\mathbf{X}$ in Lemma 3.13, as bound (ii) 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.

Thanks to Lemma 3.16, Theorem 3.4 applies and the following inequality holds true instead of (3.5).

Proposition 3.18. *Let* $\mathbf{X}_1, ..., \mathbf{X}_n$ *i.i.d. random variables, and a VC-class* $\mathcal{A}$ *with VC-dimension* $V_{\mathcal{A}}$*. Recall that* $f(\mathbf{X}_1, \dots, \mathbf{X}_n) = \sup_{A \in \mathcal{A}} \left|\mathbb{P}(\mathbf{X} \in A) - \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A}\right|$*, see* (3.2)*. Then, for* $\delta > 0$*, with probability higher than* $1 - \delta$*:*

$$f(\mathbf{X}_1, \dots, \mathbf{X}_n) \leq \mathbb{E}f(\mathbf{X}_1, \dots, \mathbf{X}_n) + \frac{2}{3n}\log\frac{1}{\delta} + 2\sqrt{\frac{q}{2n}\log\frac{1}{\delta}} \tag{3.7}$$

Proof. From Lemma 3.16 and Theorem 3.4, we have

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

or equivalently

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

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

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

Using $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ if $a,b\ge 0$, 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 $1-\delta$

$$f(\mathbf{X}_1,\ldots,\mathbf{X}_n)\le\mathbb{E}f(\mathbf{X}_1,\ldots,\mathbf{X}_n)+\frac{2}{3n}\log\frac{1}{\delta}+2\sqrt{\frac{q}{2n}\log\frac{1}{\delta}}.$$

□

Remark 3.19. (EXPECTATION BOUND) By classical arguments (see the proof of Lemma 3.13 above), $\mathbb{E}f(\mathbf{X}_1,\ldots,\mathbf{X}_n)\le q_n:=\mathbb{E}\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\left(\mathbb{1}_{\mathbf{X}_i'\in A}-\mathbb{1}_{\mathbf{X}_i\in A}\right)\right|$. Using Massart's finite class Lemma, see Massart (2000), to show that $q_n\le\sqrt{q}\sqrt{\frac{2V_\mathcal{A}\log(en/V_\mathcal{A})}{n}}$ yields

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

with probability at least $1-\delta$.

Contributions detailed in Chapter 7 use the following corollary of Proposition 3.18, see Theorem 7.1.

Corollary 3.20. *Let $\mathbf{X}_1,\ldots,\mathbf{X}_n$ i.i.d. realizations of a* r.v. $\mathbf{X}$ *and 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 such that for all $0<\delta<1$, with probability at least $1-\delta$,*

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

Proof. The inequality follows from Proposition 3.18 combined with the following Lemma which slightly improves Lemma 3.13.

Lemma 3.21. *Let $\mathbf{X}_1,...,\mathbf{X}_n$ 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:*

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

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

□

Proof of Lemma 3.21. 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 define $i.i.d.$ 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 $\sum_{i=1}^{n} \sigma_i \mathbb{1}_{\mathbf{X}_i \in A} \overset{d}{=} \sum_{i=1}^{\kappa} \sigma_i \mathbb{1}_{\mathbf{Y}_i \in A}$, where $\kappa \sim Bin(n,p)$ independent of the $\mathbf{Y}_i$'s. Thus,

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

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. □

CHAPTER 4
Extreme Value Theory

Abstract In this chapter, we provide a concise background on Extreme Value Theory (EVT). The tools needed to approach chapters 7 and 8 are introduced.

There are many books introducing extreme value theory, like Leadbetter et al. (1983), Resnick (1987), Coles et al. (2001), Beirlant et al. (2006), De Haan & Ferreira (2007), and Resnick (2007). Our favorites are Resnick (2007) for its comprehensiveness while remaining accessible, and Coles et al. (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 et al. (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 extreme 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.* '$\mathbf{x} \le \mathbf{z}$' means '$x_j \le z_j$ for all $1 \le j \le d$' and for any real number T, '$\mathbf{x} \le T$' means '$x_j \le T$ for all $1 \le j \le d$'. 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 $Y_1, \dots, Y_n$, $Y_{(1)} \le \ldots \le Y_{(n)}$ 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 $X_1,\ \ldots, X_n$ 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

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

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

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

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: $F^n(x) \to 0$ 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 that there exist 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*

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

*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 distributions (up to a re-scaling $x' = \frac{x-b}{a}$ which can be removed by changing a_n and b_n):*

$$\begin{array}{lll} \textit{Gumbel:} & G(x) = \exp\left(-e^{-x}\right) & \textit{for } x \in (-\infty, +\infty), \\ \textit{Fréchet:} & G(x) = \exp\left(-x^{-\alpha}\right) & \textit{if } x > 0 \textit{ and } G(x) = 0 \textit{ otherwise}, \\ \textit{Weibull:} & G(x) = \exp\left(-(-x)^{\alpha}\right) & \textit{if } x < 0 \textit{ and } G(x) = 1 \textit{ otherwise}, \end{array}$$

with $\alpha > 0$.

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

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

fig_source/GEV.png

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

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); Girard (2004); Boucheron & Thomas (2015), provides an estimate of the tail parameter γ. A special case of extreme quantile estimation is when some covariate information is recorded simultaneously with the quantity of interest. The extreme quantile thus depends on the covariate and is called conditional extreme quantile (Beirlant & Goegebeur, 2004; Chernozhukov, 2005; Gardes & Girard, 2008; Gardes et al., 2010; Girard & Jacob, 2008; Daouia et al., 2011, 2013).

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 ($\gamma = 0$).*

- *If the X_i's are standard Fréchet ($F(x) = \exp(-1/z)$), letting $an = n$ 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 ($\gamma = 1$).*

- *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 ($\gamma = -1$).*

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

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

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 $n \to \infty$ 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

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

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 $i = 1,\ \dots,\ n$. 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

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

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

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

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:

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

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 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

$$t\,\mathbb{P}\left(\mathbf{V} \in tA\right) \xrightarrow[t\to\infty]{} \mu(A). \tag{4.6}$$

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 $\mathbf{v} = (v_1, ..., v_d) \in \mathbb{R}^d$, set

$$\begin{cases} R(\mathbf{v}) := \|\mathbf{v}\|_\infty \ = \ \max\limits_{i=1}^{d} v_i, \\ \Theta(\mathbf{v}) := \left(\dfrac{v_1}{R(\mathbf{v})}, ..., \dfrac{v_d}{R(\mathbf{v})}\right) \in S_\infty^{d-1}, \end{cases} \tag{4.7}$$

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 $B \subset S_\infty^{d-1}$,

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

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

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

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 $\mathbf{x}^{-1} = (x_1^{-1}, \ldots, x_d^1)$. 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,

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

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,

$$t\, \mathbb{P}\left(\mathbf{U} \in t^{-1}A\right) \xrightarrow[t\to\infty]{} \Lambda(A) = \mu(A^{-1}), \tag{4.11}$$

  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 $\mathbf{x} \in [\mathbf{0}, \boldsymbol{\infty}], \mathbf{x} \neq \boldsymbol{\infty}$ by

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

  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 $G_j(x_j),\ 1 \le j \le d$,

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

In Chapter 7, 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 8 generalizes the study in Chapter 7 to the angular measure.

CHAPTER 5
Background on classical Anomaly Detection algorithms

Abstract In this chapter, we review some very classical anomaly detection algorithms used in the benchmarks of chapters 10 and 9. We also introduce the reader to the scikit-learn library used for illustrative examples, and 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 5.2.3) and Local Outlier Factor (Section 5.2.2) algorithms, as well as a participation to the scikit-learn maintenance and pull requests review.

5.1 What is Anomaly Detection?

Anomaly Detection 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. We also call them sometimes *inliers*, while abnormal data are called *outliers*. Classical parametric techniques, like those developed by Barnett & Lewis (1994) and 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 (Breunig et al., 2000; Schölkopf et al., 2001; Steinwart et al., 2005; Scott & Nowak, 2006; Vert & Vert, 2006), on dimensionality reduction (Shyu et al., 2003; Aggarwal & Yu, 2001) or on decision trees (Liu et al., 2008; Désir et al., 2012; Shi & Horvath, 2012). One may refer to Hodge & Austin (2004); Chandola et al. (2009); Patcha & Park (2007); Markou & Singh (2003) for 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/normal. Instead, they return a real valued function, termed a *scoring function*, defining a pre-order/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/ranking 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 **novelty detection** framework (also called *one-class classification* or *semi-supervised* anomaly detection), 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 (also called *outlier detection*), no assumption is made on the data which consist in unlabeled normal and abnormal instances. In general, a method from the novelty detection 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.

Let us also mention the so-called *semi-supervised novelty detection* (Blanchard et al., 2010; Smola et al., 2009) framework which is closely linked to the *PU learning framework* (Denis et al., 2005; Liu et al., 2002; Mordelet & Vert, 2014; du Plessis et al., 2015). Semi-supervised novelty detection consists in learning from negative and unsupervised examples, while PU learning consists in learning from positive (P) and unlabeled (U) examples. These hybrid approaches assume that both an unlabeled sample and a sample from one class are available.

In this thesis, we basically place ourselves in the novelty detection framework, although some benchmarks are also done on (unlabeled) training data polluted by outliers, namely in the unsupervised framework.

5.2 Three efficient Anomaly Detection Algorithms

5.2.1 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 5.1 and 5.2 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).

TABLE 5.1: SVM vs. OCSVM (hard-margin separation)

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

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

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

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 $n \to \infty$ (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.

TABLE 5.2: SVM vs. OCSVM (ν-soft margin separation)

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

(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 $O(n^3 d)$ where n is the number of samples and d the dimension of the input space. However, one can often expect a complexity of $O(n^2 d)$, 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.

5.2.2 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.

5.2.3 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. The scoring function is based on this averaged depth. Random partitioning produces noticeable shorter paths for anomalies, see figures 5.2 and 5.3. Moreover, the average depth of a sample over the forest seems to converge to some limits, the latter being different whether the sample is or not an anomaly. Hence, when a forest of random trees collectively produces shorter path lengths for particular samples, they are highly likely to be anomalies.

5.3 Examples through scikit-learn

This section provides examples on the anomaly detection algorithms presented above through the scikit-learn python library.

As mentioned in the introduction, contribution of this thesis includes the implemention 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)). 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.

5.3.1 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 favorable 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 that the user can easily construct 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 developers. 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 optionally 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.

```python
# 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 5.2.3 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 $LocalOutlierFactor()$.

[1] https://github.com/scikit-learn

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)
```

5.3.2 LOF examples

The use of LOF algorithm is illustrated in the code example below, returning Figure 5.1.

FIGURE 5.1: 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()
```

5.3.3 Isolation Forest examples

The Isolation Forest strategy is illustrated in the code example below returning Figure 5.4.

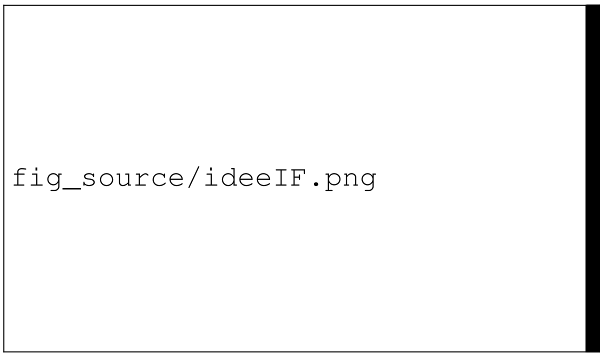


FIGURE 5.2: Anomalies are isolated more quickly

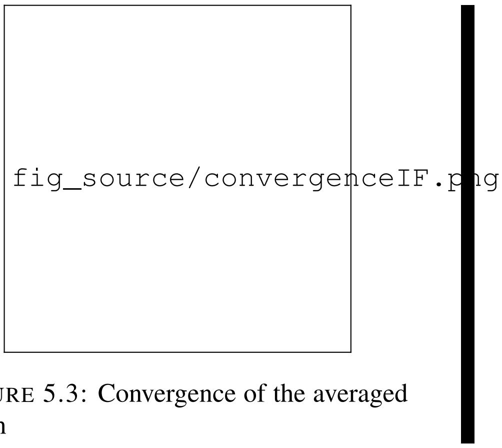


FIGURE 5.3: Convergence of the averaged depth

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

An example using IsolationForest for anomaly detection.

"""

import numpy as np
```

fig_source/iforest.png

FIGURE 5.4: Isolation Forest example

```
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()
```

5.3.4 Comparison examples

As a conclusion, Figures 5.5, 5.6 and 5.7 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 5.5: Gaussian normal data with one single mode

fig_source/ADcomparison2.png

FIGURE 5.6: Gaussian normal data with two modes

fig_source/ADcomparison3.png

FIGURE 5.7: Gaussian normal data with two strongly separate modes

```python
"""
==========================================
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()
```

PART II

An Excess-Mass based Performance Criterion

CHAPTER 6
On Anomaly Ranking and Excess-Mass Curves

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 $X_1, \ \ldots, \ X_n$ with a nearly optimal EM-curve is proposed and is analyzed from a statistical perspective.

Note: The material of this chapter is based on previous work published in Goix et al. (2015c).

6.1 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 enable 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 $X_1, \ \ldots, \ X_n$, in absence of any parametric assumptions on the underlying probability distribution describing the 'normal' regime. The most straightforward manner to define a pre-order on the feature space $\mathcal{X}$ is to transport the natural order on the real half-line through a measurable *scoring function* $s : \mathcal{X} \to \mathbb{R}_+$: 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 pre-order 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: a function having exactly the same level sets as the density can be very far from the latter using 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 non-decreasing 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 Clémençon & 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 computed.

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 Clémençon & 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 6.3 introduces the notion of EM curve and that of optimal EM curve. Estimation in the compact support case is covered by Section 6.4, extension to distributions with non compact support and control of the model bias are tackled in Section 6.5. A simulation study is performed in Section 6.6. All proofs are deferred to the last section 6.7.

6.2 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 $s : \mathbb{R}^d \to \mathbb{R}_+$ integrable *w.r.t.* Lebesgue measure. Let $s \in \mathcal{S}$. As defined in Clémençon & Jakubowicz (2013); Clémençon & 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

$$\begin{aligned} &\alpha_s(t) = \mathbb{P}(s(X) \geq t), \\ &\lambda_s(t) = \text{Leb}(\{x \in \mathbb{R}^d, s(x) \geq t\}) \end{aligned} \tag{6.1}$$

and H^{-1} denotes the pseudo-inverse of any cdf $H : \mathbb{R} \to (0,1)$. 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

$$\min_{\Gamma \ borelian} \text{Leb}(\Gamma) \text{ subject to } \mathbb{P}(X \in \Gamma) \geq \alpha. \tag{6.2}$$

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

$\mathbf{A_1}$ *The density f is bounded, i.e.* $||f(X)||_\infty < +\infty$.
$\mathbf{A_2}$ *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 (6.2) has a unique solution Γ_α^* of mass α exactly, referred to as *minimum volume set* (see Section 1.3.1):

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

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 $\widehat{\Gamma}^*_\alpha$ can be defined as the solution of (6.2) 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 approximations 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 Clémençon & Jakubowicz (2013), a method consisting in preliminarily estimating a collection of minimum volume sets related to target masses $0 < \alpha_1 < \ldots < \alpha_K < 1$ forming a subdivision of $(0,1)$ based on training data so as to build a scoring function

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

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 6.3 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 6.3 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 (6.2), 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.

6.3 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 (6.2): for $t > 0$,

$$\max_{\Omega\ borelian} \left\{\mathbb{P}(X\in\Omega) - t\mathrm{Leb}(\Omega)\right\}. \tag{6.3}$$

We denote by Ω_t^* any solution of this problem. As shall be seen in the subsequent analysis (see Proposition 6.6 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 (6.3) 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 6.1. (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 $t > \|f\|_\infty := \sup_{x \in \mathbb{R}^d} |f(x)|$.

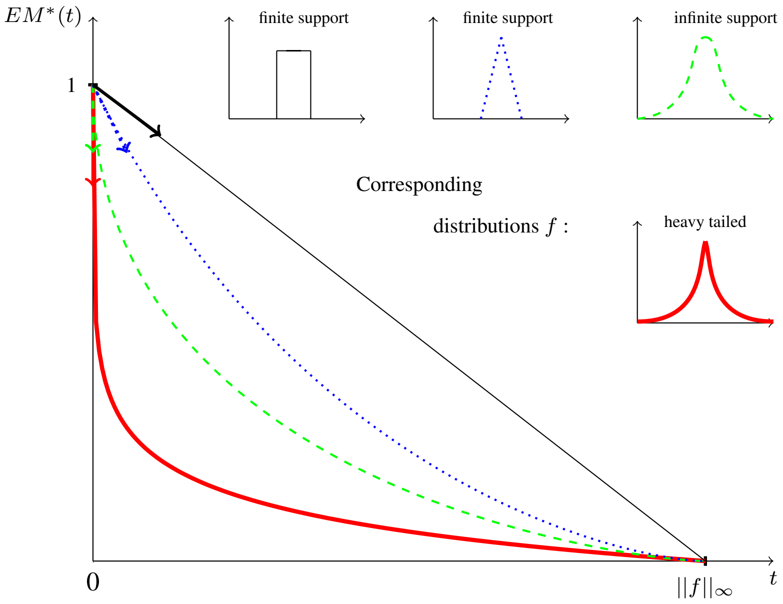


FIGURE 6.1: EM curves depending on densities

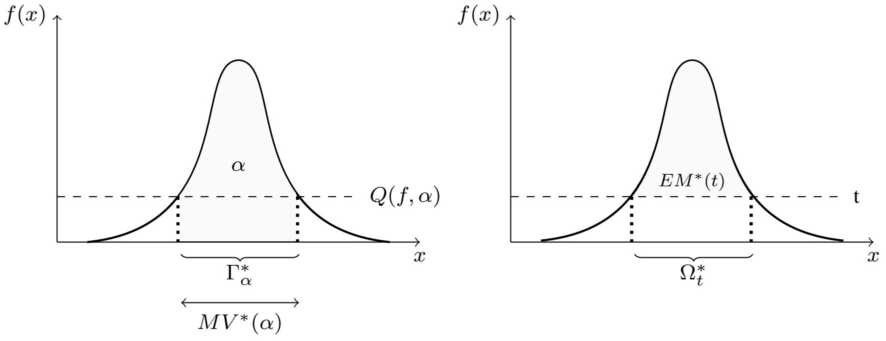


FIGURE 6.2: Comparison between $MV^*(\alpha)$ and $EM^*(t)$

Lemma 6.2. (ON EXISTENCE AND UNIQUENESS) *For any subset Ω_t^* solution of* (6.3)*, we have*

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

and the sets $\{x, f(x) > t\}$ and $\{x, f(x) \geq t\}$ are both solutions of (6.3)*. In addition, under assumption $\mathbf{A_2}$, 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 (6.1).

Proposition 6.3. (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 6.4. (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

$$EM_s : t \in [0, \infty[\mapsto \sup_{A \in \{(\Omega_{s,l})_{l>0}\}} \mathbb{P}(X \in A) - t\text{Leb}(A), \tag{6.4}$$

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 6.1 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

$$s(x) = \int_{t=0}^{+\infty} \mathbb{1}_{x \in \Omega_t^*} a(t) dt, \tag{6.5}$$

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

Proposition 6.5. *(*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 for any $\epsilon > 0$,

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

where $\{s > u\}\Delta_\epsilon\{f > t\} := \{f > t + \epsilon\} \setminus \{s > u\} \bigsqcup \{s > u\} \setminus \{f > t - \epsilon\}$ 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 $T : \mathbb{R}_+ \to \mathbb{R}_+$. 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 point-wise 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 $X_1, \ldots, X_n$ be an i.i.d. sample with common distribution $F(dx)$ and set $\widehat{\alpha}_s(t) = (1/n)\sum_{i=1}^n \mathbb{1}_{s(X_i) \geq t}$. The empirical EM curve of s is then defined as

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

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

6.4 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 Clémençon & 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 (6.3) over a subclass $\mathcal{G}$ of borelian subsets. In order to analyze the accuracy of this approach, we introduce the following additional assumptions.

$\mathbf{A_3}$ *All minimum volume sets belong to $\mathcal{G}$:*

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

$\mathbf{A_4}$ *The Rademacher average*

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

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 $\mathbf{A_4}$ is very general and is fulfilled in particular when $\mathcal{G}$ is of finite VC dimension, see Koltchinskii (2006), whereas the zero bias assumption $\mathbf{A_3}$ is in contrast very restrictive. It will be relaxed in Section 6.5.

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 $n \geq 1$:

$$\mathbb{P}\left(\left\{\sup_{G\in\mathcal{G}} (|P(G) - P_n(G)| - \Phi_n(\delta)) > 0\right\}\right) \leq \delta, \tag{6.6}$$

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 $X_1, \ \ldots, \ X_n$. For $t \geq 0$, define also the signed measures:

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

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 $0 < t_K < t_{K-1} < \ldots < t_1$. For k in $\{1, \ \ldots, \ K\}$, 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 (6.3) rather than solutions of (6.2) (corresponding to a different parametrization of the thresholds).

Proposition 6.6. (MONOTONICITY) *For any k in $\{1, \ \ldots, \ K\}$, 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 6.3, 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 6.2).

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

$$s_K(x) := \sum_{k=1}^{K} a_k \mathbb{1}_{x\in\hat{\Omega}_{t_k}}\ , \tag{6.7}$$

where $a_k > 0$ for every $k \in \{1, \ \ldots, \ K\}$. Notice that the scoring function (6.7) can be seen as a Riemann sum approximation of (6.5) 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 $K = K_n \to \infty$ as $n \to +\infty$. We assume in addition that $\sum_{k=1}^{\infty} a_k < \infty$.

Remark 6.7. (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 6.8. (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 6.9. (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 6.10. (COMPACT SUPPORT CASE) *Assume that conditions* $\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$ *and* $\mathbf{A_4}$ *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 $\sup_{1\leq k\leq K}(t_k - t_{k+1}) = \mathcal{O}(1/\sqrt{n})$, setting $t_{K+1} = 0$. 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:*

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

Remark 6.11. (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 6.10 (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 6.5.1).

Lemma 6.12. *Suppose that assumptions* $\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$ *and* $\mathbf{A_4}$ *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 6.7, and is a combination on the two following results, the second one being a straightforward consequence of the derivative property of EM^* (Proposition 6.3):

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

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

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

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

□

6.5 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 $\mathbf{A_3}$.

6.5.1 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 6.14). See Figure 6.3 as an illustration and a comparison with the MV formulation as used as a way to recover empirical minimum volume set $\hat{\Gamma}_\alpha$.

Algorithm 3 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

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

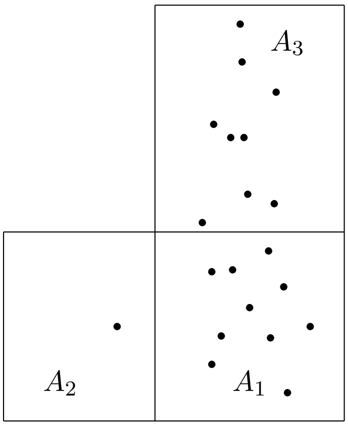
The resulting piece-wise constant scoring function is

$$s_N(x) = \sum_{k=1}^{N} (t_k - t_{k+1}) \mathbb{1}_{x \in \hat{\Omega}_{t_k}} \ . \tag{6.8}$$

The main argument to extend the above results to the case where $suppf$ is not bounded is given in Lemma 6.12 in Section 6.7. The meshgrid (t_k) must be chosen in an adaptive way, in a data-driven fashion. Let $h : \mathbb{R}_+^* \to \mathbb{R}_+$ be a decreasing function such that $\lim_{t \to 0} h(t) = +\infty$. 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 $t_1 > t_2 > \ldots > t_N > t_{N+1} = 0$, as well as $\hat{\Omega}_{t_1},\ \dots,\ \hat{\Omega}_{t_N}$, through

$$\begin{aligned} t_{k+1} &= t_k - (\sqrt{n})^{-1} \frac{1}{h(t_{k+1})} && (6.9) \\ \hat{\Omega}_{t_k} &= \arg\max_{\Omega \in \mathcal{G}} H_{n,t_k}(\Omega), && (6.10) \end{aligned}$$

with the property that $\hat{\Omega}_{t_{k+1}} \supset \hat{\Omega}_{t_k}$. As pointed out in Remark 6.9, 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



$n_1, n_2, n_3 = 10, 9, 1$

FIGURE 6.3: 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 $A_1 \cup A_3 \cup A_2$.

defined by (6.8) such that by virtue of Lemma 6.12 (see technical details in Section 6.7), 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 $t \to 0$, we can assume that $\lambda(t)/h(t) \leq B$ for t in $]0, t_1]$ since λ is decreasing, and we obtain:

$$\sup_{t\in]t_N,t_1]} |EM^*(t) - EM_{s_N}(t)| \ \leq\ \left(A + \sqrt{2\log(1/\delta)}\right) \frac{1}{\sqrt{n}}\ . \tag{6.11}$$

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, (6.10) yields, for $k \geq 2$,

$$t_k = \frac{t_1}{(1+\frac{1}{\sqrt{n}})^{k-1}}\ . \tag{6.12}$$

Remark 6.13. Theorem 6.10 holds true with the t_k's defined as in (6.12) – even if the condition $\sup_{1\leq k\leq K}(t_k - t_{k+1}) = \mathcal{O}(1/\sqrt{n})$ is not respected – as soon as $t_K = \mathcal{O}(1/\sqrt{n})$.

Theorem 6.14. (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* (6.12), Ω_{t_k} *by (6.9), and* s_N *(6.8). 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 $o_N(1) = 1 - EM^*(t_N)$ *represents 'how much* f *is heavy tailed'. In addition,* $s_N(x)$ *converges to* $s_\infty(x) := \sum_{k=1}^{\infty}(t_{k+1} - t_k)\mathbb{1}_{\hat{\Omega}_{t_{k+1}}}$ *as* $N \to \infty$ *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 6.14 (Sketch of). The first assertion is a consequence of (6.11) 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 6.7, $EM_{s_N} \leq EM_{s_\infty}$. For a detailed proof, see Section 6.7. □

6.5.2 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.

$\tilde{\mathbf{A}}_\mathbf{3}$ *There exists a countable sub-collection 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 piece-wise 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 6.10 and 6.14 can be established without assumption $\mathbf{A_3}$, as soon as $\tilde{\mathbf{A}}_\mathbf{3}$ 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 6.14:

Theorem 6.15. *(*Biased empirical clusters*) Suppose that assumptions* $\mathbf{A_1}$*,* $\mathbf{A_2}$*,* $\tilde{\mathbf{A}}_\mathbf{3}$*,* $\mathbf{A_4}$ *hold true, let* $t_1 > 0$ *and for* $k \geq 2$ *consider* t_k *defined by (6.12),* Ω_{t_k} *by (6.9), and* s_N *by (6.8). 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 6.16. (Hyper-cubes) 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 $\|f - f_{F_l}\|_{L^1} \to 0$ as $l \to 0$. For instance, the class F_l could be formed by disjoint hyper-cubes of side length l.

Proof of Theorem 6.15 (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 6.17. *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 6.7 for details), one shows that:

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

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. □

6.6 Simulation examples

Algorithm 3 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 $F_l = \{F^l_i\}_{i \in \mathbb{Z}^2}$ and $F^l_i = [li_1, li_1 + 1] \times [li_2, li_2 + 1]$ 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 $l \to 0$ (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 hyper-cubes, 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 6.4 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 6.5 reveals an ‘over-fitting’ 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 over-fitting is experienced because of the high local density of the observations. Recalling the notation $EM^*_{\mathcal{G}}(t) = \max_{\Omega \in \mathcal{G}} H_t(\Omega) \leq EM^*(t) = \max_{\Omega \; meas.} H_t(\Omega)$ so that the bias of our model is $EM^* - EM^*_{\mathcal{G}}$, Figure 6.6 illustrates the variations of the bias with the wealth of our model characterized by l the width of the partition by hyper-cubes. 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 over-fit the extreme data, 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 hyper-cubes 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.

FIGURE 6.4: Optimal and realized EM curves

FIGURE 6.5: Zoom near 0

FIGURE 6.6: EM_G for different l

Conclusion

Prolongating the contribution of Clémençon & Jakubowicz (2013), this chapter provides an alternative view (respectively, an other parametrization) of the anomaly scoring problem, leading 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 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.

Illustrations

Note that the scoring function we built in Algorithm 3 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 6.7. Note that the silhouette does not focus on local properties of the density, but only on its induced pre-order (level sets).

fig_source/scoring3D.png

FIGURE 6.7: density and scoring functions

6.7 Detailed Proofs

Proof of Proposition 6.3

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 $h \to 0$, and expressing $EM^{*'} = \alpha'(t) - t\lambda'(t) - \lambda(t)$, it suffices to show that $\alpha'(t) - t\lambda'(t) = 0$, namely $\lim_{h \to 0} A(h) - t\, B(h) =$

0. Yet, we have

$$A(h) - t\,B(h) \;=\; \frac{1}{h}\int_{t \le f \le t+h} f - t \;\le\; \frac{1}{h}\int_{t \le f \le t+h} h \;=\; \mathrm{Leb}(t \le f \le t+h) \to 0$$

because f has no flat part.

Proof of Lemma 6.2:

On the one hand, for every Ω measurable,

$$\begin{aligned}\mathbb{P}(X \in \Omega) - t\,\mathrm{Leb}(\Omega) &= \int_\Omega (f(x)-t)dx \\ &\le \int_{\Omega \cap \{f \ge t\}} (f(x)-t)dx \\ &\le \int_{\{f \ge t\}} (f(x)-t)dx \\ &= \mathbb{P}(f(X) \ge t) - t\,\mathrm{Leb}(\{f \ge t\}).\end{aligned}$$

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

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

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

so that

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

i.e

$$\mathbb{P}(X \in \Omega) - t\,\mathrm{Leb}(\Omega) \;<\; \mathbb{P}(f(X) \ge t) - t\,\mathrm{Leb}(\{x, f(x) \ge t\})$$

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

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

$$\int_{\Omega_t^* \cap \{f < t-\epsilon\}} f - t \le -\epsilon.\mathrm{Leb}(\Omega_t^* \cap \{f < t-\epsilon\}) < 0.$$

It follows that

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

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

Proof of Proposition 6.5

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,

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

so we obtain:

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

The other inequality comes from the fact that

$$\begin{aligned} &\int_{\{f>t\}\backslash\{s>u\}}(f(x)-t)dx + \int_{\{s>u\}\backslash\{f>t\}}(t-f(x))dx \\ &\qquad\qquad \geq \int_{\{f>t+\epsilon\}\backslash\{s>u\}}(f(x)-t)dx + \int_{\{s>u\}\backslash\{f>t+\epsilon\}}(f(x)-t)dx \\ &\qquad\qquad \geq \epsilon\text{Leb}(\{f>t+\epsilon\}\setminus\{s>u\}) + \epsilon\text{Leb}(\{s>u\}\setminus\{f>t+\epsilon\}) \end{aligned}$$

To prove the third point, note that:

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

Yet,

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

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

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:

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

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

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

Proof of Proposition 6.6

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

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

It follows that

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

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 $k + m \leq K$:

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

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 6.10

We shall use the following lemma:

Lemma 6.18. *With probability at least* $1-\delta$*, for* $k \in \{1, ..., K\}$*,*

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

Proof of Lemma 6.18:

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:

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

On the other hand, using (6.6), with probability at least $1-\delta$, for every $G \in \mathcal{G}$, $|\mathbb{P}(G) - \mathbb{P}_n(G)| \le \Phi_n(\delta)$. Hence, with probability at least $1-\delta$, for all $\Omega \in \mathcal{G}$:

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

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

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

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

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

The following Lemma is a consequence of the derivative property of EM^* (Proposition 6.3) .

Lemma 6.19. *Let* k *in* $\{1, ..., K-1\}$*. Then for every* t *in* $]t_{k+1}, t_k]$*,*

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

Combined with Lemma 6.18 and the fact that EM_{s_K} is non-increasing, and writing

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

this result leads to:

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

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

which gives Lemma 6.12 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 6.19 to $k = K$, namely the inequality holds true for t in $]t_{K+1}, t_K] =]0, t_K]$ as soon as we let $\lambda(t_{K+1}) :=$ Leb$(suppf)$. Indeed the compactness of $suppf$ implies that $\lambda(t) \to$ Leb$(suppf)$ as $t \to 0$. Observing that Lemma 6.18 already contains the case $k = K$, this leads to, for k in $\{0, ..., K\}$ and $t \in\ ]t_{k+1}, t_k]$, $|EM^*(t) - EM_{s_K}(t)| \le 2\Phi_n(\delta) + \lambda(t_{k+1})(t_k - t_{k+1})$. Therefore, λ being a decreasing function bounded by Leb$(suppf)$, we obtain the following: with probability at least $1-\delta$, we have for all t in $]0,t_1]$,

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

Proof of Theorem 6.14

The first part of this theorem is a consequence of (6.11) 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 6.18.
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 6.7 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 6.17

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:

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

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

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

PART III

Accuracy on Extreme Regions

CHAPTER 7
Learning the dependence structure of rare events: a non-asymptotic study

Abstract This chapter presents the details relative to the introducing section 1.4.1. 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).

7.1 Introduction

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 7.2 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 modeling 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 (2007) 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 7.2. In Section 7.3, 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. The main result of this contribution, a non-asymptotic bound on the convergence of the empirical STDF, is derived in Section 7.4. Section 7.5 concludes.

7.2 Background on the stable tail dependence function

In the multivariate case, it is mathematically very convenient to decompose the joint distribution of $\mathbf{X} = (X^1, \ldots, X^d)$ 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 $\mathbf{U} = (U^1, \ldots, U^d)$. Knowledge of the F_j's and of the joint distribution of $\mathbf{U}$ allows to recover that of $\mathbf{X}$, since $\mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d) = \mathbb{P}(U^1 \geq 1 - F_1(x_1), \ldots, U^d \geq 1 - F_d(x_d))$. 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

$$\lim_{t \to 0} t^{-1} \mathbb{P}\left[U^1 \leq t\, x_1 \text{ or } \ldots \text{ or } U^d \leq t\, x_d\right] = \Lambda[\mathbf{x}, \infty]^c := l(\mathbf{x}) \,. \qquad (x_j \in [0, \infty], \mathbf{x} \neq \infty) \tag{7.1}$$

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 (7.1) is called the *stable tail dependence function*. In the remainder of this chapter, the only assumption is the existence of a limit in (7.1), *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 $\mathbf{x} = (x_1, \ldots, x_d)$ (*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,

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

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 $l(.)$ 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.

7.3 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 7.1. *Let* $\mathbf{X}_1, \dots, \mathbf{X}_n$ *i.i.d. realizations of a* r.v. $\mathbf{X}$ *and 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 *such that for all* $0 < \delta < 1$*, with probability at least* $1 - \delta$*,*

$$\sup_{A\in\mathcal{A}} \left| \mathbb{P}\big[\mathbf{X} \in A\big] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A} \right| \;\leq\; C\left[\sqrt{p}\sqrt{\frac{V_{\mathcal{A}}}{n}\log\frac{1}{\delta}} + \frac{1}{n}\log\frac{1}{\delta}\right]. \tag{7.2}$$

Proof. See Chapter 3, Corollary 3.20. □

Remark 7.2. (COMPARISON WITH EXISTING BOUNDS) The following re-normalized VC-inequality is due to Vapnik and Chervonenkis (see Vapnik & Chervonenkis (1974), Anthony

& Shawe-Taylor (1993) or Bousquet et al. (2004), Thm 7),

$$\sup_{A\in\mathcal{A}}\left|\frac{\mathbb{P}(\mathbf{X}\in A)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{X}_i\in A}}{\sqrt{\mathbb{P}(\mathbf{X}\in A)}}\right| \;\leq\; 2\sqrt{\frac{\log S_{\mathcal{A}}(2n)+\log\frac{4}{\delta}}{n}}\,. \tag{7.3}$$

$S_{\mathcal{A}}(n)$ is the shattering coefficient (or growth function) associated with class $\mathcal{A}$. (7.3) holds under the same conditions as Theorem 7.1, and allows to derive a bound similar to (7.2), 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 (7.3) 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 7.3. (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 7.4. (INTERPRETATION) Inequality (7.2) 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 (7.2) by p, so that the left hand side becomes a supremum of conditional probabilities upon belonging to the union class $\mathbb{A}$, $\{\mathbb{P}(\mathbf{X}\in A\big|\mathbf{X}\in\mathbb{A})\}_{A\in\mathbb{A}}$. 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 (7.2), and can thus be viewed as the effective sample size.

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 $g:\mathcal{X}\to\{-1,1\}$, 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 $\{(\mathbf{X}_1,Y_1),\ldots,(\mathbf{X}_n,Y_n)\}$. 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

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

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 $\|\mathbf{X}_{(1)}\| \geq \ldots \geq \|\mathbf{X}_{(n)}\|$ are the order statistics of $\|\mathbf{X}\|$. Then as an application of Theorem 7.1 with $\mathcal{A} = \{(\mathbf{x}, y), g(\mathbf{x}) \neq y, \|\mathbf{x}\| > t_\alpha\},\ g \in \mathcal{G}$, we have :

$$\sup_{g \in \mathcal{G}}\left|L_{\alpha,n}(g) - L_\alpha(g)\right| \leq C\left[\sqrt{\frac{V_{\mathcal{G}}}{n\alpha}\log\frac{1}{\delta}} + \frac{1}{n\alpha}\log\frac{1}{\delta}\right]. \tag{7.4}$$

We refer to the remark 7.5 below for details. Again the obtained rate by empirical risk minimization meets our expectations (see remark 7.4), 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, $\alpha = \alpha_n \to 0$. 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 $n \to +\infty$ 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 $\widetilde{L}_Q(g) := \mathbb{P}\left(Y \neq g(\mathbf{X}), \mathbf{X} \in Q\right)$ and $\widetilde{L}_{Q,n}(g) := \mathbb{P}_n\left(Y \neq g(\mathbf{X}), \mathbf{X} \in Q\right)$; 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 $1 - \alpha$.

Remark 7.5. To obtain the bound in (7.4), the following easy to show inequality is needed before applying Theorem 7.1 :

$$\begin{aligned}\sup_{g\in\mathcal{G}}\left|L_{\alpha,n}(g) - L_\alpha(g)\right| \;\leq\; \frac{1}{\alpha}\Bigg[&\sup_{g\in\mathcal{G}}\left|\mathbb{P}\left(Y \neq g(\mathbf{X}),\ \|\mathbf{X}\|\ > t_\alpha\right) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{\{Y_i \neq g(\mathbf{X}_i),\ \|\mathbf{X}_i\| > t_\alpha\}}\right| \\ &+ \left|\mathbb{P}\left(\|\mathbf{X}\|\ > t_\alpha\right) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{\{\|\mathbf{X}_i\| > t_\alpha\}}\right| \;+\; \frac{1}{n}\Bigg]. \ \blacksquare\end{aligned}$$

Note that the final objective would be to bound the quantity $|L_\alpha(g_n) - L_\alpha(g_\alpha^*)|$, where g_α^* is a Bayes classifier for the problem at stake, *i.e.* a solution of the conditional risk minimization problem $\inf_{\{g \text{ meas.}\}} L_\alpha(g)$, and g_n a solution of $\inf_{g\in\mathcal{G}} L_{n,\alpha}(g)$. Such a bound involves the maximal deviation in (7.4) as well as 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

$$\frac{1}{\alpha}\inf_{g\in\mathcal{G}}\mathbb{E}\left[|2\eta(\mathbf{X}) - 1|\mathbb{1}_{g(\mathbf{X}) \neq g^*(\mathbf{X})}\mathbb{1}_{\|\mathbf{X}\| \geq t_\alpha}\right],$$

to be compared with the standard bias

$$\inf_{g\in\mathcal{G}} \mathbb{E}\left[|2\eta(\mathbf{X})-1|\mathbb{1}_{g(\mathbf{X})\neq g^*(\mathbf{X})}\right].$$

7.4 A bound on the STDF

Let us place ourselves in the multivariate extreme framework introduced in Section 7.1: 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 (7.1) and the margins F_j are supposed to be unknown. The definition of l may be recast as

$$l(\mathbf{x}):=\lim_{t\to 0} t^{-1}\tilde{F}(t\mathbf{x}) \tag{7.5}$$

with $\tilde{F}(\mathbf{x})=(1-F)\big((1-F_1)^{\leftarrow}(x_1),\dots,(1-F_d)^{\leftarrow}(x_d)\big)$. 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, $\tilde{F}(\mathbf{x})=\mathbb{P}\Big(\bigcup_{j=1}^d\{U^j\leq x_j\}\Big)=\mathbb{P}(\mathbf{U}\in[\mathbf{x},\infty[^c)$.

Let $k=k(n)$ be a sequence of positive integers such that $k\to\infty$ and $k=o(n)$ as $n\to\infty$. A natural estimator of l is its empirical version defined as follows, see Huang (1992), Qi (1997), Drees & Huang (1998), Einmahl et al. (2006):

$$l_n(\mathbf{x})=\frac{1}{k}\sum_{i=1}^n \mathbb{1}_{\{X_i^1\geq X^1_{(n-\lfloor kx_1\rfloor+1)}\text{ or } \dots \text{ or } X_i^d\geq X^d_{(n-\lfloor kx_d\rfloor+1)}\}}, \tag{7.6}$$

The expression is indeed suggested by the definition of l in (7.5), 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 non-parametric estimator of l, see Huang (1992), Drees & Huang (1998) and De Haan & Ferreira (2007) 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 7.1 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

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

This is a VC-class of VC-dimension d, as proved in Devroye et al. (1996), Theorem 13.8, for its complementary class $\big\{[\mathbf{x},\infty[,\ \mathbf{x}>0\big\}$. 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 $C(\mathbf{x}) = \mathbb{P}(Z \in\,\cdot\,)$. As a direct consequence of Theorem 7.1 the following inequality holds true with probability at least $1-\delta$,

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

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

$$\sup_{0\le\mathbf{x}\le T} \frac{n}{k}\left|C_n(\frac{k}{n}[\mathbf{x},\infty[^c) - C(\frac{k}{n}[\mathbf{x},\infty[^c)\right| \;\le\; Cd\sqrt{\frac{T}{k}\log\frac{1}{\delta}}\;. \tag{7.7}$$

Inequality (7.7) 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 $k(n) \to \infty$. To keep the notation uncluttered, we often drop the dependence in n and simply write k instead of $k(n)$.

Theorem 7.6. *Let T be a positive number such that $T \ge \frac{7}{2}(\frac{\log d}{k}+1)$, and δ such that $\delta \ge e^{-k}$. Then there is an absolute constant C such that for each $n > 0$, with probability at least $1-\delta$:*

$$\sup_{0\le\mathbf{x}\le T} |l_n(\mathbf{x}) - l(\mathbf{x})| \;\le\; Cd\sqrt{\frac{T}{k}\log\frac{d+3}{\delta}} \;+\; \sup_{0\le\mathbf{x}\le 2T}\left|\frac{n}{k}\tilde{F}(\frac{k}{n}\mathbf{x}) - l(\mathbf{x})\right| \tag{7.8}$$

The second term on the right hand side of (7.8) is a bias term which depends on the discrepancy between the left hand side and the limit in (7.1) or (7.5) 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 7.6 follows the same lines as in Qi (1997). For one-dimensional random variables $Y_1,\dots,Y_n$, let us denote by $Y_{(1)} \le \ldots \le Y_{(n)}$ their order statistics. Define then the empirical version $\tilde{F}_n$ of $\tilde{F}$ (introduced in (7.5)) as

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

so that $\frac{n}{k}\tilde{F}_n(\frac{k}{n}\mathbf{x}) \;=\; \frac{1}{k}\sum_{i=1}^n \mathbb{1}_{\{U_i^1 \le \frac{k}{n}x_1 \text{ or ... or } U_i^d \le \frac{k}{n}x_d\}}$. 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 7.7 (Link between l_n and $\tilde{F}_n$). *The empirical version of $\tilde{F}$ and that of l are related* via

$$l_n(\mathbf{x}) \;=\; \frac{n}{k}\tilde{F}_n(U^1_{(\lfloor kx_1\rfloor)},\;\cdots\;,U^d_{(\lfloor kx_d\rfloor)}).$$

Proof. Consider the definition of l_n in (7.6), and note that for $j = 1, \dots, d$,

$$\begin{aligned} X_i^j \geq X^j_{(n - \lfloor kx_i \rfloor + 1)} &\Leftrightarrow rank(X_i^j) \geq n - \lfloor kx_j \rfloor + 1 \\ &\Leftrightarrow rank(F_j(X_i^j)) \geq n - \lfloor kx_j \rfloor + 1 \\ &\Leftrightarrow rank(1 - F_j(X_i^j)) \leq \lfloor kx_j \rfloor \\ &\Leftrightarrow U_i^j \leq U^j_{(\lfloor kx_j \rfloor)}, \end{aligned}$$

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

Lemma 7.8 (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:*

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

Proof. Notice that

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

and apply inequality (7.7). □

Lemma 7.9 (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 $1 - \delta$,*

$$\forall\, 1 \leq j \leq d, \quad \frac{n}{k} U^j_{(\lfloor kT \rfloor)} \;\leq\; 2T\,, \tag{7.9}$$

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

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

Proof. Notice that $\sup_{[0,T]} \frac{n}{k} U^j_{(\lfloor k\,\cdot \rfloor)} = \frac{n}{k} U^j_{(\lfloor kT \rfloor)}$ 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

$$\frac{n}{k} U^j_{(\lfloor kT \rfloor)} \leq 2T \;\Leftrightarrow\; \Gamma_n\Big(\frac{k}{n} 2T\Big) \geq \frac{\lfloor kT \rfloor}{n}$$

so that

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

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

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

and thus

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

as required in (7.9). Yet,

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

where $\Theta_j(y) = \frac{n}{k}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{U_i^j \le \frac{k}{n}y\}} - \mathbb{P}\left[U_1^j \le \frac{k}{n}y\right]\right|$. Then, by (7.9), with probability greater than $1-\delta$,

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

and from (7.7), each term $\sup_{0\le y\le 2T}\Theta_j(y)$ is bounded by $C\sqrt{\frac{T}{k}\log\frac{1}{\delta}}$ (with probability $1-\delta$). In the end, with probability greater than $1-(d+1)\delta$:

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

which is the desired inequality □

We may now proceed with the proof of Theorem 7.6. 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 (7.5) it is easy to prove by sub-dividing $[0,T]^d$ (see Qi (1997) p.174 for details) that

$$\sup_{0\le\mathbf{x}\le T}\left|\frac{1}{t}\tilde{F}(t\mathbf{x}) - l(\mathbf{x})\right| \to 0 \quad \text{as} \quad \mathrm{t}\to 0\,. \tag{7.10}$$

Using Lemma 7.7, we can write :

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

Now, by (7.9) we have with probability greater than $1-\delta$:

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

and by Lemma 7.8,

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

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

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

by virtue of (7.10). Concerning $\Upsilon(n)$, we have :

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

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

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

so that by Lemma 7.9, with probability greater than $1-(d+1)\delta$:

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

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

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

7.5 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 7.3 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.

CHAPTER 8
Sparse Representation of Multivariate Extremes

Abstract This chapter presents the details relative to the introducing Section 1.4.2. 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 (non-parametric) 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 behaviors, to be confronted with new (possibly abnormal) extreme observations. Illustrative experimental results provide strong empirical evidence of the relevance of our approach.

Note: The material of this chapter is based on previous work under review available in Goix et al. (2016b). Part of this work have been published in Goix et al. (2016c) and Goix et al. (2015a).

8.1 Introduction

8.1.1 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 $\mathbf{Y} = (Y_1, ..., Y_d)$ 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 8.2), $C\Phi(A) \simeq \mathbb{P}(\|\mathbf{Y}\|^{-1}\mathbf{Y} \in A \mid \|\mathbf{Y}\| > r)$, 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 $\{\max_{1 \leq i \leq k} x_i = 1 \; ; \; x_i > 0 \; (i \leq k) \; ; \; x_{k+1} = \ldots = x_d = 0\}$ indicates that the components $Y_1, \ldots, Y_k$ 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. (2006).

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 (2014) 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 (2007).

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 complex 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 (2014)). 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 looses 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 7, 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 (8.6), Section 8.2). 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 7 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.

8.1.2 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 8.2. A non-parametric estimator of the subfaces' mass is introduced in Section 8.3, 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 8.4, 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 8.5. Technical details are deferred at the end of this chapter, Section 8.7.

8.2 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.* $\mathbf{X} = (X^1, \ \ldots, \ X^d)$ with distribution $\mathbf{F}(\mathbf{x}) := \mathbb{P}(X_1 \le x_1, \ldots, X_d \le x_d)$, namely $\mathbf{F} \in \mathbf{DA}(\mathbf{G})$ it stipulates the existence of two sequences $\{\mathbf{a}_n, n \ge 1\}$ and $\{\mathbf{b}_n, n \ge 1\}$ in $\mathbb{R}^d$, the $\mathbf{a}_n$'s being positive, and a non-degenerate distribution function $\mathbf{G}$ such that

$$\lim_{n \to \infty} n\, \mathbb{P}\left(\frac{X^1 - b_n^1}{a_n^1} \ \ge \ x_1 \text{ or } \ldots \text{ or } \frac{X^d - b_n^d}{a_n^d} \ \ge \ x_d \right) = -\log \mathbf{G}(\mathbf{x}) \qquad (8.1)$$

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 $\mathbf{V} = (V^1, \ \ldots, \ V^d)$, Assumption (8.1) implies the existence of a limit measure μ on $[0, \infty]^d \setminus \{\mathbf{0}\}$ such that

$$n\, \mathbb{P}\left(\frac{V^1}{n} \ \ge \ v_1 \text{ or } \cdots \text{ or } \frac{V^d}{n} \ \ge \ v_d \right) \xrightarrow[n \to \infty]{} \mu\left([\mathbf{0}, \mathbf{v}]^c \right), \qquad (8.2)$$

where $[\mathbf{0}, \mathbf{v}] := [0, \ v_1] \times \cdots \times [0, \ v_d]$. The dependence structure of the limit $\mathbf{G}$ in (8.1) can then be expressed by means of the so-termed *exponent measure* μ:

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

The measure μ should be viewed, up to a 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

$$t\,\mathbb{P}\left(\mathbf{V} \in tA\right) \xrightarrow[t\to\infty]{} \mu(A). \tag{8.3}$$

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 $\mathbf{v} = (v_1, ..., v_d) \in \mathbb{R}^d$, set

$$\begin{cases} R(\mathbf{v}) := \|\mathbf{v}\|_\infty \;=\; \max\limits_{i=1}^{d} v_i, \\ \Theta(\mathbf{v}) := \left(\dfrac{v_1}{R(\mathbf{v})}, ..., \dfrac{v_d}{R(\mathbf{v})}\right) \in S_\infty^{d-1}, \end{cases} \tag{8.4}$$

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 $B \subset S_\infty^{d-1}$,

$$\mu\{\mathbf{v} \,:\, R(\mathbf{v}) > z, \Theta(\mathbf{v}) \in B\} = z^{-1}\Phi(B)\;. \tag{8.5}$$

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

$$\mu\big([\mathbf{0}, \mathbf{x}^{-1}]^c\big) = \int_{\boldsymbol{\theta}\in S_\infty^{d-1}} \max_j \theta_j x_j \;\mathrm{d}\Phi(\boldsymbol{\theta}). \tag{8.6}$$

Recall that here and beyond, operators on vectors are understood component-wise, so that $\mathbf{x}^{-1} = (x_1^{-1}, \ldots, x_d^1)$. 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,

$$\mathbb{P}(\Theta \in A \mid R > r) = \frac{r\mathbb{P}(\Theta \in A, R > r)}{r\mathbb{P}(R > r)} \;\xrightarrow[r\to\infty]{}\; \frac{\Phi(A)}{\Phi(S_\infty^{d-1})}. \tag{8.7}$$

8.2.1 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 (8.1)). 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 $\alpha \subset \{1, \;\ldots, \; d\}$, consider the truncated cone (see Fig. 8.1)

$$\mathcal{C}_\alpha = \{\mathbf{v} \geq 0,\; \|\mathbf{v}\|_\infty \geq 1,\; v_j > 0 \;\text{ for } j \in \alpha,\; v_j = 0 \;\text{ for } j \notin \alpha\}. \tag{8.8}$$

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 $\mathbb{R}_+^d \setminus [\mathbf{0}, \mathbf{1}]$, one may naturally decompose the exponent measure as

$$\mu = \sum_{\emptyset \neq \alpha \subset \{1,\dots,d\}} \mu_\alpha, \tag{8.9}$$

where each component μ_α is concentrated on the untruncated cone corresponding to $\mathcal{C}_\alpha$. 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 $\mathcal{C}_\alpha$ 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 $\mathcal{C}_\alpha$ 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 $t \to +\infty$. 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*

$$R_\alpha^\epsilon \ = \{\mathbf{v} \geq 0,\ \|\mathbf{v}\|_\infty \geq 1,\ v_j > \epsilon \ \text{ for } j \in \alpha,\ v_j \leq \epsilon \ \text{ for } j \notin \alpha\}, \tag{8.10}$$

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 $\mathcal{C}_\alpha$ is the same as nonzero mass on R_α^ϵ for ϵ arbitrarily small.

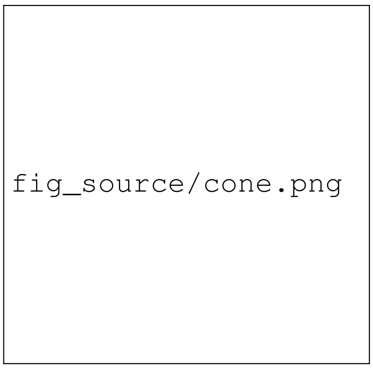


FIGURE 8.1: Truncated cones in 3D

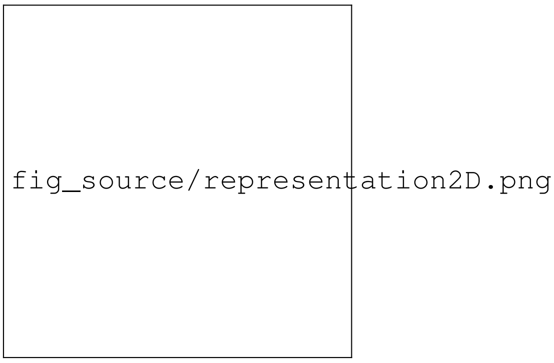


FIGURE 8.2: Truncated ϵ-rectangles in 2D

Lemma 8.1. *For any non empty index subset $\emptyset \neq \alpha \subset \{1,\dots,d\}$, the exponent measure of $\mathcal{C}_\alpha$ is*

$$\mu(\mathcal{C}_\alpha) = \lim_{\epsilon \to 0} \mu(R_\alpha^\epsilon).$$

Proof. First consider the case $\alpha = \{1, \dots, d\}$. Then R^ϵ_α's forms an increasing sequence of sets as ϵ decreases and $\mathcal{C}_\alpha = R^0_\alpha = \cup_{\epsilon>0,\epsilon\in\mathbb{Q}} R^\epsilon_\alpha$. 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

$$O^\epsilon_\alpha = \{\mathbf{x} \in \mathbb{R}^d_+ \ : \forall j \in \alpha : x_j > \epsilon\},$$
$$N^\epsilon_\alpha = \{\mathbf{x} \in \mathbb{R}^d_+ \ : \forall j \in \alpha : x_j > \epsilon, \exists j \notin \alpha : x_j > \epsilon\},$$

so that $N^\epsilon_\alpha \subset O^\epsilon_\alpha$ and $R^\epsilon_\alpha = O^\epsilon_\alpha \setminus N^\epsilon_\alpha$. Observe also that $\mathcal{C}_\alpha = O^0_\alpha \setminus N^0_\alpha$. Thus, $\mu(R^\epsilon_\alpha) = \mu(O^\epsilon_\alpha) - \mu(N^\epsilon_\alpha)$, and $\mu(\mathcal{C}_\alpha) = \mu(O^0_\alpha) - \mu(N^0_\alpha)$, so that it is sufficient to show that

$$\mu(N^0_\alpha) = \lim_{\epsilon\to 0} \mu(N^\epsilon_\alpha), \quad \text{and} \quad \mu(O^0_\alpha) = \lim_{\epsilon\to 0} \mu(O^\epsilon_\alpha).$$

Notice that the N^ϵ_α's and the O^ϵ_α's form two increasing sequences of sets (when ϵ decreases), and that $N^0_\alpha = \bigcup_{\epsilon>0,\epsilon\in\mathbb{Q}} N^\epsilon_\alpha$, $O^0_\alpha = \bigcup_{\epsilon>0,\epsilon\in\mathbb{Q}} O^\epsilon_\alpha$. This proves the desired result. □

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 8.1, and the discussion above, ϵ may be chosen such that R^ϵ_α is a continuity set of μ, while $\mu(R^\epsilon_\alpha)$ is arbitrarily close to $\mu(\mathcal{C}_\alpha)$. Then, using the characterization (8.3) of μ, the following asymptotic identity holds true:

$$\begin{aligned}\lim_{t\to\infty} t\mathbb{P}\left(\|\mathbf{V}\|_\infty \geq t, V^j > \epsilon t \ \ (j \in \alpha), V^j \leq \epsilon t \ \ (j \notin \alpha)\right) &= \mu(R^\epsilon_\alpha) \\ &\simeq \mu(\mathcal{C}_\alpha).\end{aligned} \tag{8.11}$$

Remark 8.2. In terms of conditional probabilities, denoting $R = \|T(\mathbf{X})\|$, where T is the standardization map $\mathbf{X} \mapsto \mathbf{V}$, we have

$$\mathbb{P}(T(\mathbf{X}) \in rR^\epsilon_\alpha \mid R > r) = \frac{r\mathbb{P}(\mathbf{V} \in rR^\epsilon_\alpha)}{r\mathbb{P}(\mathbf{V} \in r([\mathbf{0},\mathbf{1}]^c)} \xrightarrow[r\to\infty]{} \frac{\mu(R^\epsilon_\alpha)}{\mu([\mathbf{0},\mathbf{1}]^c)},$$

as in (8.7). In other terms,

$$\begin{aligned}\mathbb{P}\left(V^j > \epsilon r \ \ (j \in \alpha), V^j \leq \epsilon r \ \ (j \notin \alpha) \;\middle|\; \|\mathbf{V}\|_\infty \geq r\right) &\xrightarrow[r\to\infty]{} C\mu(R^\epsilon_\alpha) \\ &\simeq C\mu(\mathcal{C}_\alpha),\end{aligned} \tag{8.12}$$

where $C = 1/\Phi(S^{d-1}_\infty) = 1/\mu([\mathbf{0},\mathbf{1}]^c)$. 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 $\mu'_\alpha s$ (*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 $\mathcal{C}_\alpha$'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 8.3.1, evaluated on the ϵ-thickened rectangles R^ϵ_α. Formally, we wish to recover the

$(2^d - 1)$-dimensional unknown vector

$$\mathcal{M} = \{\mu(\mathcal{C}_\alpha) : \emptyset \neq \alpha \subset \{1, \ldots, d\}\} \tag{8.13}$$

from $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} \mathbf{F}$ and build an estimator $\widehat{\mathcal{M}}$ such that

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

is small with large probability. In view of Lemma 8.1, (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 8.3.1 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.

8.2.2 Regularity Assumptions

Beyond the existence of the limit measure μ (*i.e.* multivariate regular variation of $\mathbf{V}$'s distribution, see (8.2)), and thus, existence of an angular measure Φ (see (8.5)), 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 $\{1, \ldots, d\}$, one denotes by $\mathrm{d}x_\alpha$ the Lebesgue measure on $\mathcal{C}_\alpha$ and write $\mathrm{d}x_\alpha = \mathrm{d}x_{i_1} \ldots \mathrm{d}x_{i_k}$, when $\alpha = \{i_1, \ldots, i_k\}$. For convenience, we also write $\mathrm{d}x_{\alpha \setminus i}$ instead of $\mathrm{d}x_{\alpha \setminus \{i\}}$.

Assumption 2. Each component μ_α of (8.9) is absolutely continuous w.r.t. Lebesgue measure $\mathrm{d}x_\alpha$ on $\mathcal{C}_\alpha$.

Assumption 2 has a very convenient consequence regarding Φ: the fact that the exponent measure μ spreads no mass on subsets of the form $\{\mathbf{x} : \|\mathbf{x}\|_\infty \geq 1, x_{i_1} = \cdots = x_{i_r} \neq 0\}$ with $r \geq 2$, implies that the spectral measure Φ spreads no mass on edges $\{\mathbf{x} : \|\mathbf{x}\|_\infty = 1,\ x_{i_1} = \cdots = x_{i_r} = 1\}$ with $r \geq 2$. This is summarized by the following result.

Lemma 8.3. *Under Assumption 2, the following assertions holds true.*

- Φ *is concentrated on the (disjoint) edges*

$$\begin{aligned} \Omega_{\alpha,i_0} = \{\mathbf{x} : \|\mathbf{x}\|_\infty = 1,\ x_{i_0} = 1,\ & 0 < x_i < 1 \quad \textit{for } i \in \alpha \setminus \{i_0\} \\ & x_i = 0 \qquad \textit{for } i \notin \alpha \qquad \} \end{aligned} \tag{8.14}$$

for $i_0 \in \alpha$, $\emptyset \neq \alpha \subset \{1, \ldots, d\}$.

- *The restriction* Φ_{α,i_0} *of* Φ *to* Ω_{α,i_0} *is absolutely continuous* w.r.t. *the Lebesgue measure* $\mathrm{d}x_{\alpha\setminus i_0}$ *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 $\int_D \mathrm{d}x_{\alpha\setminus i_0} = 0$. Then the induced truncated cone $\tilde{D} = \{\mathbf{v} : \ \|\mathbf{v}\|_\infty \geq 1, \mathbf{v}/\|\mathbf{v}\|_\infty \in D\}$ satisfies $\int_{\tilde{D}} \mathrm{d}x_\alpha = 0$ and belongs to $\mathcal{C}_\alpha$. Thus, by virtue of Assumption 2, $\Phi_{\alpha,i_0}(D) = \Phi_\alpha(D) = \mu_\alpha(\tilde{D}) = 0$. □

It follows from Lemma 8.3 that the angular measure Φ decomposes as $\Phi = \sum_\alpha \sum_{i_0\in\alpha} \Phi_{\alpha,i_0}$ and that there exist densities $\frac{\mathrm{d}\Phi_{\alpha,i_0}}{\mathrm{d}x_{\alpha\smallsetminus i_0}}$, $|\alpha| \geq 2$, $i_0 \in \alpha$, such that for all $B \subset \Omega_\alpha$, $|\alpha| \geq 2$,

$$\Phi(B) \ = \ \Phi_\alpha(B) \ = \ \sum_{i_0\in\alpha} \int_{B\cap\Omega_{\alpha,i_0}} \frac{\mathrm{d}\Phi_{\alpha,i_0}}{\mathrm{d}x_{\alpha\smallsetminus i_0}}(x)\mathrm{d}x_{\alpha\setminus i_0}. \tag{8.15}$$

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

$$M_\beta = \sup_{i\in\beta} \ \sup_{x\in\Omega_{\beta,i}} \ \frac{\mathrm{d}\Phi_{\beta,i}}{\mathrm{d}x_{\beta\setminus i}}(x). \tag{8.16}$$

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 8.4. The constant M is problem dependent. However, in the case where our representation $\mathcal{M}$ defined in (8.13) 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, $M \leq \sum_{|\beta|\geq 2} M_\beta|\beta| = \sum_{|\beta|\geq 2} \Phi(\Omega_\beta) \leq \sum_\beta \Phi(\Omega_\beta) \leq \mu([\mathbf{0},\mathbf{1}]^c)$. 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 8.3, $\Omega_\alpha = \bigsqcup_{i_0\in\alpha} \Omega_{\alpha,i_0}$, each of the edges Ω_{α,i_0} being unit hypercube. Now, $\mu([\mathbf{0},\mathbf{1}]^c) \leq \mu(\{v,\ \exists j,\ v_j > 1\} \leq d\mu(\{v,\ v_1 > 1\})) \leq d$.

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 (8.13) 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 8.3 depends on the sparsity constant M.

8.3 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 8.2.2). Assumptions 2 and 3 are not necessary to prove a preliminary result on a class of rectangles (Proposition 8.8 and Corollary 8.9). However, they are required to bound the bias induced by the tolerance parameter ϵ (in Lemma 8.10, Proposition 8.11 and in the main result, Theorem 8.12).

8.3.1 A natural empirical version of the exponent measure mu

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

$$\widehat{\mathbf{V}}_i = \left((1 - \widehat{F}_j(X_i^j))^{-1}\right)_{1 \le j \le d},$$

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),

$$T(\mathbf{x}) = \left(\frac{1}{1 - F_j(x^j)}\right)_{1 \le j \le d} \quad \text{and} \quad \widehat{T}(\mathbf{x}) = \left(\frac{1}{1 - \widehat{F}_j(x^j)}\right)_{1 \le j \le d}. \tag{8.17}$$

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, $t\,\mathbb{P}\left(\mathbf{V} \in tA\right) \to \mu(A)$ as $t \to \infty$, see (8.3), a natural empirical version of μ is defined as

$$\mu_n(A) \;=\; \frac{n}{k}\widehat{\mathbb{P}}_n(\frac{n}{k}A) \;=\; \frac{1}{k}\sum_{i=1}^n \mathbf{1}_{\{\widehat{\mathbf{V}}_i \in \frac{n}{k}A\}}\,. \tag{8.18}$$

Here and throughout, we place ourselves in the asymptotic setting stipulating that $k = k(n) > 0$ is such that $k \to \infty$ and $k = o(n)$ as $n \to \infty$. 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).

8.3.2 Accounting for the non asymptotic nature of data: epsilon-thickening.

Since the cones $\mathcal{C}_\alpha$ 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 8.2.1 and Lemma 8.1, it is natural to introduce a tolerance parameter $\epsilon > 0$ and to approximate the asymptotic mass of $\mathcal{C}_\alpha$ with the non-asymptotic mass of R_α^ϵ. We thus define the non-parametric estimator $\widehat{M}(\alpha)$ of $\mu(\mathcal{C}_\alpha)$ as

$$\widehat{\mathcal{M}}(\alpha) = \mu_n(R_\alpha^\epsilon), \qquad \emptyset \neq \alpha \subset \{1, \dots, d\}. \tag{8.19}$$

Evaluating $\widehat{\mathcal{M}}(\alpha)$ boils down (see (8.18)) to counting points in $(n/k)\,R_\alpha^\epsilon$, as illustrated in Figure 8.3. 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 $\mathcal{C}_\alpha$ constitute a summary representation of the dependence structure. This representation is sparse as soon as the $\mu_n(R_\alpha^\epsilon)$ are positive only for a few groups of features α (compared to the total number of groups or sub-cones, 2^d

FIGURE 8.3: Estimation procedure

namely). It is 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 8.2,

$$\mathcal{M}(\alpha) = \lim_{r\to\infty} \mu([\mathbf{0},\mathbf{1}]^c)\ \mathbb{P}(T(\mathbf{X}) \in rR_\alpha^\epsilon \mid \|T(\mathbf{X})\| \geq r). \tag{8.20}$$

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 8.12. 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:

$$\begin{aligned} ||\widehat{\mathcal{M}} - \mathcal{M}||_\infty &= \max_\alpha |\mu_n(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)| \\ &\leq \max_\alpha |\mu - \mu_n|(R_\alpha^\epsilon) + \max_\alpha |\mu(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)| \,. \end{aligned} \tag{8.21}$$

Here and beyond, to keep the notation uncluttered, we simply denotes 'α' for 'α non empty subset of $\{1,\ \ldots,\ d\}$'. The main steps of the argument leading to Theorem 8.12 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 8.3.3), and deduce an uniform bound on $|\mu_n - \mu|(R_\alpha^\epsilon)$ (Subsection 8.3.4). Finally, using the regularity assumptions (Assumption 2 and Assumption 3), bound the difference $|\mu(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)|$ (Subsection 8.3.5).

8.3.3 Preliminaries: uniform approximation over a VC-class of rectangles

This subsection builds on the theory developed in Chapter 7, where a non-asymptotic bound is stated on the estimation of the stable tail dependence function (defined in (4.12)). The STDF l is related to the class of sets of the form $[\mathbf{0}, \mathbf{v}]^c$ (or $[\mathbf{u}, \infty]^c$ depending on which standardization is used), and an equivalent definition is

$$l(\mathbf{x}) := \lim_{t\to\infty} t\tilde{F}(t^{-1}\mathbf{x}) = \mu([\mathbf{0}, \mathbf{x}^{-1}]^c) \tag{8.22}$$

with $\tilde{F}(\mathbf{x}) = (1-F)\big((1-F_1)^{\leftarrow}(x_1), \dots, (1-F_d)^{\leftarrow}(x_d)\big)$. 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,

$$\tilde{F}(\mathbf{x}) = \mathbb{P}\Big(\bigcup_{j=1}^{d}\{U^j < x_j\}\Big) = \mathbb{P}(\mathbf{U} \in [\mathbf{x}, \infty[^c) = \mathbb{P}(\mathbf{V} \in [\mathbf{0}, \mathbf{x}^{-1}]^c). \tag{8.23}$$

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):

$$l_n(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{X_i^1 \geq X^1_{(n-\lfloor kx_1\rfloor+1)} \text{ or } \dots \text{ or } X_i^d \geq X^d_{(n-\lfloor kx_d\rfloor+1)}\}}\,. \tag{8.24}$$

The expression is indeed suggested by the definition of l in (8.22), 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 8.5. *Consider the rank transformed variables $\widehat{\mathbf{U}}_i = (\widehat{\mathbf{V}}_i)^{-1} = (1 - \widehat{F}_j(X_i^j))_{1\leq j\leq d}$ for $i = 1, \dots, n$. Then, for $(i,j) \in \{1,\dots,n\} \times \{1,\dots,d\}$, with probability one,*

$$\widehat{U}_i^j \leq \frac{k}{n}x_j^{-1} \Leftrightarrow \widehat{V}_i^j \geq \frac{n}{k}x_j \Leftrightarrow X_i^j \geq X^j_{(n-\lfloor kx_j^{-1}\rfloor+1)} \Leftrightarrow U_i^j \leq U^j_{(\lfloor kx_j^{-1}\rfloor)}\,.$$

The proof of Lemma 8.5 is standard and is provided in Section 8.7 for completeness. By Lemma 8.5, the following alternative expression of $l_n(\mathbf{x})$ holds true:

$$l_n(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{U_i^1 \,\leq\, U^1_{(\lfloor kx_1\rfloor)} \text{ or } \dots \text{ or } U_i^d \,\leq\, U^d_{(\lfloor kx_d\rfloor)}\}} = \mu_n\left([\mathbf{0}, \mathbf{x}^{-1}]^c\right). \tag{8.25}$$

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 (2007) 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

$$\begin{aligned} R(\mathbf{x}, \mathbf{z}, \alpha, \beta) \; = \; \Big\{ & \mathbf{y} \in [0, \infty]^d, \;\; y_j \geq x_j \;\; \text{for } j \in \alpha, \\ & \qquad\qquad y_j < z_j \;\; \text{for } j \in \beta \;\; \Big\}, \quad \mathbf{x}, \mathbf{z} \in [0, \infty]^d. \end{aligned} \tag{8.26}$$

Thus,

$$\mu_n \left(R(\mathbf{x}, \mathbf{z}, \alpha, \beta) \right) \; = \; \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}_{\{\widehat{V}_i^j \, \geq \, \frac{n}{k} x_j \text{ for } j \in \alpha \text{ and } \widehat{V}_i^j \, < \, \frac{n}{k} x_j \text{ for } j \in \beta\}} \, .$$

Then, define the functional $g_{\alpha,\beta}$ (which plays the same role as the STDF) as follows: for $\mathbf{x} \in [0, \infty]^d \setminus \{\boldsymbol{\infty}\}$, $\mathbf{z} \in [0, \infty]^d$, $\alpha \subset \{1, \ldots, d\} \setminus \emptyset$ and $\beta \subset \{1, \ldots, d\}$, let

$$g_{\alpha,\beta}(\mathbf{x}, \mathbf{z}) \;\; = \;\; \lim_{t \to \infty} t \tilde{F}_{\alpha,\beta}(t^{-1}\mathbf{x}, t^{-1}\mathbf{z}), \;\; \text{with} \tag{8.27}$$

$$\tilde{F}_{\alpha,\beta}(\mathbf{x}, \mathbf{z}) \;\; = \;\; \mathbb{P}\Big[\big\{ U^j \leq x_j \;\; \text{for } j \in \alpha \big\} \; \bigcap \; \big\{ U^j > z_j \;\; \text{for } j \in \beta \big\} \Big]. \tag{8.28}$$

Notice that $\tilde{F}_{\alpha,\beta}(\mathbf{x}, \mathbf{z})$ is an extension of the non-asymptotic approximation $\tilde{F}$ in (8.22). By (8.27) and (8.28), we have

$$\begin{aligned} g_{\alpha,\beta}(\mathbf{x}, \mathbf{z}) &= \lim_{t \to \infty} t \mathbb{P} \Big[\big\{ U^j \leq t^{-1} x_j \;\; \text{for } j \in \alpha \big\} \; \bigcap \; \big\{ U^j > t^{-1} z_j \;\; \text{for } j \in \beta \big\} \Big] \\ &= \lim_{t \to \infty} t \mathbb{P} \left[\mathbf{V} \in t R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \; \beta) \right] \; , \end{aligned}$$

so that using (8.3),

$$g_{\alpha,\beta}(\mathbf{x}, \mathbf{z}) = \mu([R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \; \beta)]). \tag{8.29}$$

The following lemma makes the relation between $g_{\alpha,\beta}$ and the angular measure Φ explicit. Its proof is given in Section 8.7.

Lemma 8.6. *The function $g_{\alpha,\beta}$ can be represented as follows:*

$$g_{\alpha,\beta}(\mathbf{x}, \mathbf{z}) = \int_{S^{d-1}} \left(\bigwedge_{j \in \alpha} w_j x_j - \bigvee_{j \in \beta} w_j z_j \right)_+ \Phi(\mathrm{d}\mathbf{w}) \; ,$$

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 8.7. Lemma 8.6 shows that the functional $g_{\alpha,\beta}$, which plays the same role as a 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 (8.24)) by

$$g_{n,\alpha,\beta}(\mathbf{x}, \mathbf{z}) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}_{\{X_i^j \geq X_{(n - \lfloor k x_j \rfloor + 1)}^j \;\; \text{for } j \in \alpha \;\; \text{and} \;\; X_i^j < X_{(n - \lfloor k x_j \rfloor + 1)}^j \;\; \text{for } j \in \beta\}} \, . \tag{8.30}$$

As it is the case for the empirical STDF (see (8.25)), $g_{n,\alpha,\beta}$ has an alternative expression

$$\begin{aligned} g_{n,\alpha,\beta}(\mathbf{x},\mathbf{z}) &= \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{U_i^j \leq U^j_{([kx_j])} \text{ for } j\in\alpha \text{ and } U_i^j > U^j_{([kx_j])} \text{ for } j\in\beta\}} \\ &= \mu_n\left(R(\mathbf{x}^{-1},\mathbf{z}^{-1},\alpha,\ \beta)\right), \end{aligned} \tag{8.31}$$

where the last equality comes from the equivalence $\widehat{V}_i^j \geq \frac{n}{k}x_j \Leftrightarrow U_i^j \leq U^j_{(\lfloor kx_j^{-1}\rfloor)}$ (Lemma 8.5) 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 (8.18).

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 $\{1,\ldots,d\}\setminus\emptyset$ and β subset of $\{1,\ldots,d\}$'. 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 8.8. *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 $1-\delta$,*

$$\begin{aligned} \max_{\alpha,\beta}\ \sup_{0\leq\mathbf{x},\mathbf{z}\leq T} |g_{n,\alpha,\beta}(\mathbf{x},\mathbf{z}) - g_{\alpha,\beta}(\mathbf{x},\mathbf{z})| \ &\leq\ Cd\sqrt{\frac{2T}{k}\log\frac{d+3}{\delta}} \\ &+ \max_{\alpha,\beta}\ \sup_{0\leq\mathbf{x},\mathbf{z}\leq 2T}\left|\frac{n}{k}\tilde{F}_{\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z}) - g_{\alpha,\beta}(\mathbf{x},\mathbf{z})\right|. \end{aligned} \tag{8.32}$$

The second term on the right hand side of the inequality is an asymptotic bias term which goes to 0 as $n\to\infty$ (see Remark 8.24).

The proof follows the same lines as that of Theorem 6 in Goix et al. (2015b) and is detailed in Section 8.7. 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.

8.3.4 Bounding empirical deviations over thickened rectangles

The aim of this subsection is to bound $|\mu_n-\mu|(R^\epsilon_\alpha)$ 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 $\alpha \subset \{1,\ldots,d\}$. In the remainder of the chapter, $\bar{\alpha}$ denotes the complementary set of α in $\{1,\ldots,d\}$. Notice that directly from their definitions (8.10) and (8.26), R_α^ϵ and $R(\mathbf{x},\mathbf{z},\alpha,\beta)$ are linked by:

$$R_\alpha^\epsilon = R(\boldsymbol{\epsilon},\boldsymbol{\epsilon},\alpha,\bar{\alpha}) \cap [\mathbf{0},\mathbf{1}]^c = R(\boldsymbol{\epsilon},\boldsymbol{\epsilon},\alpha,\bar{\alpha}) \setminus R(\boldsymbol{\epsilon},\tilde{\boldsymbol{\epsilon}},\alpha,\{1,\ldots,d\})$$

where $\tilde{\boldsymbol{\epsilon}}$ is defined by $\tilde{\boldsymbol{\epsilon}}_j = \mathbb{1}_{j\in\alpha} + \epsilon\mathbb{1}_{j\notin\alpha}$ for all $j \in \{1,\ldots,d\}$. Indeed, we have: $R(\boldsymbol{\epsilon},\boldsymbol{\epsilon},\alpha,\bar{\alpha}) \cap [\mathbf{0},\mathbf{1}] = R(\boldsymbol{\epsilon},\tilde{\boldsymbol{\epsilon}},\alpha,\{1,\ldots,d\})$. As a result, for $\epsilon < 1$,

$$\sup_{\epsilon\le\mathbf{x},\mathbf{z}} |\mu_n - \mu|\,(R_\alpha^\epsilon) \le 2 \sup_{\epsilon\le\mathbf{x},\mathbf{z}} |\mu_n - \mu|\,(R(\mathbf{x},\ \mathbf{z},\ \alpha,\ \bar{\alpha}))\,.$$

On the other hand, from (8.31) and (8.29) we have

$$\sup_{\epsilon\le\mathbf{x},\mathbf{z}} |\mu_n - \mu|\,(R(\mathbf{x},\ \mathbf{z},\ \alpha,\ \bar{\alpha})) \ = \ \sup_{0\le\mathbf{x},\mathbf{z}\le\epsilon^{-1}} |g_{n,\alpha,\bar{\alpha}}(\mathbf{x},\mathbf{z}) - g_{\alpha,\bar{\alpha}}(\mathbf{x},\mathbf{z})|\,.$$

Then Proposition 8.8 applies with $T = 1/\epsilon$ and the following result holds true.

Corollary 8.9. *Let* $0 < \epsilon \le (\frac{7}{2}(\frac{\log d}{k}+1))^{-1}$, *and* $\delta \ge e^{-k}$. *Then there is a universal constant* C, *such that for each* $n > 0$, *with probability at least* $1-\delta$,

$$\begin{aligned}\max_\alpha \sup_{\epsilon\le\mathbf{x},\mathbf{z}} |(\mu_n-\mu)(R_\alpha^\epsilon)| \ &\le\ Cd\sqrt{\frac{1}{\epsilon k}\log\frac{d+3}{\delta}} \\ &+ \max_{\alpha,\beta}\sup_{0\le\mathbf{x},\mathbf{z}\le 2\epsilon^{-1}}\left|\frac{n}{k}\tilde{F}_{\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z}) - g_{\alpha,\beta}(\mathbf{x},\mathbf{z})\right|.\end{aligned} \tag{8.33}$$

8.3.5 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 (8.10) do not correspond to any set of angles on the sphere S_∞^{d-1}, we also define the (ϵ,ϵ')*-thickened cones*

$$\mathcal{C}_\alpha^{\epsilon,\epsilon'} \ = \{\mathbf{v}\ge 0,\ \|\mathbf{v}\|_\infty \ge 1,\ v_j > \epsilon\|\mathbf{v}\|_\infty \ \text{for } j\in\alpha,\ v_j \le \epsilon'\|\mathbf{v}\|_\infty \ \text{ for } j\notin\alpha\}, \tag{8.34}$$

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

$$\Omega_\alpha^{\epsilon,\epsilon'} \ = \ \left\{\mathbf{x}\in S_\infty^{d-1},\ x_i > \epsilon \ \text{ for } i\in\alpha\,,\ x_i \le \epsilon' \ \text{ for } i\notin\alpha\right\} = \ \mathcal{C}_\alpha^{\epsilon,\epsilon'} \cap S_\infty^{d-1}. \tag{8.35}$$

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

$$\Phi(\Omega_\alpha^{\epsilon,0}) = \mu(\mathcal{C}_\alpha^{\epsilon,0}) \le \mu(R_\alpha^\epsilon) \le \mu(\mathcal{C}_\alpha^{0,\epsilon}) = \Phi(\Omega_\alpha^{0,\epsilon}). \tag{8.36}$$

The next result (proved in Section 8.7) 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 8.10. *For every* $\emptyset \ne \alpha \subset \{1,\ldots,d\}$ *and* $0 < \epsilon,\epsilon' < 1/2$, *we have*

$$|\Phi(\Omega_\alpha^{\epsilon,\epsilon'}) - \Phi(\Omega_\alpha)| \ \le\ M|\alpha|^2\epsilon \ +\ Md\epsilon'\,.$$

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 8.11. *For every non empty set of indices $\emptyset \neq \alpha \subset \{1,\dots,d\}$ and $\epsilon > 0$,*

$$|\mu(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)| \leq Md^2\epsilon$$

8.3.6 Main result

We can now state the main result of the contribution, revealing the accuracy of the estimate (8.19).

Theorem 8.12. *There is an universal constant $C > 0$ such that for every $n,\ k,\ \epsilon,\ \delta$ 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 $1 - \delta$:*

$$\begin{aligned} \|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty \ &\leq \ Cd\left(\sqrt{\frac{1}{\epsilon k}\log\frac{d}{\delta}} + Md\epsilon\right) \\ &\quad + \ 4 \max_{\substack{\alpha \,\subset\, \{1,\dots,d\} \\ \alpha \neq \emptyset}} \ \sup_{0 \leq \mathbf{x},\mathbf{z} \leq \frac{2}{\epsilon}} \left| \frac{n}{k}\tilde{F}_{\alpha,\bar{\alpha}}(\frac{k}{n}\mathbf{x}, \frac{k}{n}\mathbf{z}) - g_{\alpha,\bar{\alpha}}(\mathbf{x},\mathbf{z}) \right|. \end{aligned}$$

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 $n \to \infty$ (see Remark 8.24). 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 (8.21). The first term $\sup_\alpha |\mu_n(R_\alpha^\epsilon) - \mu(R_\alpha^\epsilon)|$ on the right hand side of (8.21) is bounded using Corollary 8.9, while Proposition 8.11 allows to bound the second one (bias term stemming from the tolerance parameter ϵ). Introduce the notation

$$\mathrm{bias}(\alpha, n, k, \epsilon) = 4 \sup_{0 \leq \mathbf{x},\mathbf{z} \leq \frac{2}{\epsilon}} \left| \frac{n}{k}\tilde{F}_{\alpha,\bar{\alpha}}(\frac{k}{n}\mathbf{x}, \frac{k}{n}\mathbf{z}) - g_{\alpha,\bar{\alpha}}(\mathbf{x},\mathbf{z}) \right|. \tag{8.37}$$

With probability at least $1 - \delta$,

$$\begin{aligned} &\forall\, \emptyset \neq \alpha \subset \{1,\dots,d\}, \\ &|\mu_n(R_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)| \ \leq \ Cd\sqrt{\frac{1}{\epsilon k}\log\frac{d+3}{\delta}} \ + \ \mathrm{bias}(\alpha, n, k, \epsilon) + Md^2\epsilon\ . \end{aligned}$$

The upper bound stated in Theorem 8.12 follows.

Remark 8.13. (THRESHOLD ON 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.* 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:

$$\begin{aligned}
\|\widetilde{\mathcal{M}} - \mathcal{M}\|_\infty &\le \|\widetilde{\mathcal{M}} - \widehat{\mathcal{M}}\|_\infty + \|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty \\
&\le p|A|^{-1}\sum_\alpha \widehat{\mathcal{M}}(\alpha) + \|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty \\
&\le p|A|^{-1}\sum_\alpha \mathcal{M}(\alpha) + p|A|^{-1}\sum_\alpha |\widehat{\mathcal{M}}(\alpha) - \mathcal{M}(\alpha)| \\
&\qquad\qquad + \|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty \\
&\le (p+1)\|\widehat{\mathcal{M}} - \mathcal{M}\|_\infty + p|A|^{-1}\mu([0,1]^c).
\end{aligned}$$

It is outside the scope of this chapter to study optimal values for p. However, Remark 8.14 writes the estimation procedure as an optimization problem, thus exhibiting a link between thresholding and L^1-regularization.

Remark 8.14. (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 $\widehat{\mathcal{M}}(\alpha) = 1/k\sum_{i=1}^n \mathbb{1}_{\frac{k}{n}\hat{\mathbf{V}}_i \in R_\alpha^\epsilon}$. As $R_\alpha^\epsilon \subset [\mathbf{0},\mathbf{1}]^c$, we may write

$$\widehat{\mathcal{M}}(\alpha) = \Big(\frac{n}{k}\mathcal{P}_n(\frac{k}{n}\|\hat{\mathbf{V}}_1\| \ge 1)\Big)\ \Big(\frac{1}{n}\sum_{i=1}^n \frac{\mathbb{1}_{\frac{k}{n}\hat{\mathbf{V}}_i \in R_\alpha^\epsilon}\mathbb{1}_{\frac{k}{n}\|\hat{\mathbf{V}}_i\| \ge 1}}{\mathcal{P}_n(\frac{k}{n}\|\hat{\mathbf{V}}_1\| \ge 1)}\Big),$$

where the last term is the empirical expectation of $Z_{n,i}(\alpha) = \mathbb{1}_{\frac{k}{n}\hat{\mathbf{V}}_i \in R_\alpha^\epsilon}$ conditionally to the event $\{\|\frac{k}{n}\hat{\mathbf{V}}_1\| \ge 1\}$, and $\mathcal{P}_n(\frac{k}{n}\|\hat{\mathbf{V}}_1\| \ge 1) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\frac{k}{n}\|\hat{\mathbf{V}}_i\| \ge 1}$. According to Lemma 8.5, for each fixed margin j, $\hat{V}_i^j \ge \frac{n}{k}$ if, and only if $X_i^j \ge X_{(n-k+1)}^j$, which happens for k observations exactly. Thus,

$$\mathcal{P}_n(\frac{k}{n}\|\hat{\mathbf{V}}_1\| \ge 1) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\exists j, \hat{\mathbf{V}}_i^j \ge \frac{n}{k}} \in \left[\frac{k}{n}, \frac{dk}{n}\right].$$

If we define $\tilde{k} = \tilde{k}(n) \in [k, dk]$ such that $\mathcal{P}_n(\frac{k}{n}\|\hat{\mathbf{V}}_1\| \ge 1) = \frac{\tilde{k}}{n}$, we then have

$$\begin{aligned}
\widehat{\mathcal{M}}(\alpha) &= \frac{\tilde{k}}{k}\ \left(\frac{1}{n}\sum_{i=1}^n \frac{\mathbb{1}_{\frac{k}{n}\hat{\mathbf{V}}_i \in R_\alpha^\epsilon}\mathbb{1}_{\frac{k}{n}\|\hat{\mathbf{V}}_i\| \ge 1}}{\mathcal{P}_n(\frac{k}{n}\|\hat{\mathbf{V}}_1\| \ge 1)}\right) \\
&= \frac{\tilde{k}}{k}\ \operatorname*{arg\,min}_{m_\alpha > 0} \sum_{i=1}^n (Z_{n,i}(\alpha) - m_\alpha)^2 \mathbb{1}_{\frac{k}{n}\|\hat{\mathbf{V}}_i\| \ge 1},
\end{aligned}$$

Considering now the (2^d-1)-vector $\widehat{\mathcal{M}}$ and $\|.\|_{2,\alpha}$ the L^2-norm on $\mathbb{R}^{2^d-1}$, we immediately have (since $k(n)$ does not depend on α)

$$\widehat{\mathcal{M}} = \frac{\tilde{k}}{k}\operatorname*{arg\,min}_{m \in \mathbb{R}^{2^d-1}} \widehat{R_n}(m), \tag{8.38}$$

where $\widehat{R_n}(m) = \sum_{i=1}^n \|Z_{n,i} - m\|_{2,\alpha}^2 \mathbb{1}_{\frac{k}{n}\|\hat{\mathbf{V}}_i\| \geq 1}$ is the L^2-empirical risk of m, restricted to extreme observations, namely to observations $\mathbf{X}_i$ satisfying $\|\hat{\mathbf{V}}_i\| \geq \frac{n}{k}$. 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

$$R_n(m) = \mathbb{E}\left[\|Z_n - m\|_{2,\alpha}^2 \Big| \|\frac{k}{n}\mathbf{V}_1\|_\infty \geq 1\right]$$

with $Z_n := Z_{n,1}$. Then one can show (see Section 8.7) 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 $R_n(m) \to R_\infty(m) := \mathbb{E}[\|Z_\infty - m\|_{2,\alpha}^2]$, which is the asymptotic risk. Moreover, the optimization problem

$$\min_{m \in \mathbb{R}^{2^d-1}} R_\infty(m)$$

admits $m = (\frac{\mu(R_\alpha^\epsilon)}{\mu([\mathbf{0},\mathbf{1}]^c)}, \alpha \subset \{1, \dots, n\}, \alpha \neq \emptyset)$ as solution.

Considering the solution of the minimization problem (8.38), 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 penality 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 8.13.

8.4 Application to Anomaly Detection

8.4.1 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 novelty detection (training data only contain normal data) and unsupervised (training data contain unlabeled normal and abnormal data) 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.

8.4.2 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 8.2 and 8.3. 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 survival function $x \mapsto 1/x$) is convenient: Consider thus the $\mathbf{V}_i$'s and $\widehat{\mathbf{V}}_i$'s as defined in Section 8.2. One possible strategy to investigate the dependence structure of extreme events is to characterize, for each subset of features $\alpha \subset \{1, ..., d\}$, 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 (8.10). 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 (8.19)), with μ_n the empirical version of μ defined in (8.18), namely:

$$\widehat{\mathcal{M}}(\alpha) = \mu_n(R_\alpha^\epsilon) = \frac{n}{k}\widehat{\mathbb{P}}_n\left(\frac{n}{k}R_\alpha^\epsilon\right), \tag{8.39}$$

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 $k \to \infty$ and $k = o(n)$ as $n \to \infty$. The ratio

n/k plays the role of a large radial threshold r. From our standardization choice, counting points in $(n/k)\,R^\epsilon_\alpha$ boils down to selecting, for each feature $j \le d$, the 'k largest values' X^j_i 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 $(\widehat{\mathcal{M}}(\alpha))_{\alpha\subset\{1,\dots,d\}}$ associated with the cones $\mathcal{C}_\alpha$ 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 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 $\mathcal{C}_\alpha$, conditioned upon exceeding a large threshold. Indeed, for r, n and k sufficiently large, we have (Remark 8.2 and (8.20), reminding that $\mathbf{V} = T(\mathbf{X})$)

$$\widehat{\mathcal{M}}(\alpha) \simeq C\mathbb{P}(T(\mathbf{X}) \in rR^\epsilon_\alpha \mid \|T(\mathbf{X})\| \ge r).$$

Introduce an 'angular scoring function'

$$w_n(\mathbf{x}) = \sum_\alpha \widehat{\mathcal{M}}(\alpha)\mathbb{1}_{\{\widehat{T}(\mathbf{x})\in R^\epsilon_\alpha\}}. \tag{8.40}$$

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 $T(\mathbf{x}) \in R^\epsilon_\alpha$. Consider the '*directional tail region*' induced by $\mathbf{x}$, $A_\mathbf{x} = \{\mathbf{y} : T(\mathbf{y}) \in R^\epsilon_\alpha,\ \|T(\mathbf{y})\|_\infty \ge \|T(\mathbf{x})\|_\infty\}$. Then, if $\|T(\mathbf{x})\|_\infty$ is large enough, we have (using (8.5)) that

$$\begin{aligned}
\mathbb{P}\left(\mathbf{X} \in A_\mathbf{x}\right) &= \mathbb{P}\left(\mathbf{V} \in \|T(\mathbf{x})\|_\infty R^\epsilon_\alpha\right)\\
&= \mathbb{P}\left(\|\mathbf{V}\| \ge \|T(\mathbf{x})\|\right)\ \mathbb{P}\left(\mathbf{V} \in \|T(\mathbf{x})\|_\infty R^\epsilon_\alpha \mid \|\mathbf{V}\| \ge \|T(\mathbf{x})\|\right)\\
&\simeq C\,\mathbb{P}\left(\|\mathbf{V}\| \ge \|T(\mathbf{x})\|\right)\ \widehat{\mathcal{M}}(\alpha)\\
&= C\,\|\widehat{T}(\mathbf{x})\|_\infty^{-1}\,w_n(\mathbf{x}).
\end{aligned}$$

This yields the scoring function

$$s_n(\mathbf{x}) := \frac{w_n(\mathbf{x})}{\|\widehat{T}(\mathbf{x})\|_\infty}, \tag{8.41}$$

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 8.4 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* fig_source/plot_damex_level_sets.png

FIGURE 8.4: Level sets of s_n on simulated 2D data

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. (2016c), where 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 $\widehat{F}_j(X_i^j)$ (Step 1) by sorting the data (*e.g.* merge sort). The second one arises from Step 2.

Algorithm 4 DAMEX

Input: parameters $\epsilon > 0, \quad k = k(n), \quad p \geq 0.$

1. Standardize *via* marginal rank-transformation: $\widehat{\mathbf{V}}_i := \big(1/(1 - \widehat{F}_j(X_i^j))\big)_{j=1,\dots,d}$.
2. Assign to each $\widehat{\mathbf{V}}_i$ the cone R_α^ϵ it belongs to.
3. Compute $\widehat{\mathcal{M}}(\alpha)$ from (8.39) $\rightarrow$ yields: (small number of) cones with non-zero mass.
4. (Optional) Set to 0 the $\widehat{\mathcal{M}}(\alpha)$ below some small threshold defined in remark 8.13 *w.r.t. p.*$\rightarrow$ yields: (sparse) representation of the dependence structure

$$\left\{\widehat{\mathcal{M}}(\alpha) : \ \emptyset\alpha \subset \{1, \dots, d\}\right\}. \tag{8.42}$$

Output: Compute the scoring function given by (8.41),

$$s_n(\mathbf{x}) := (1/\|\widehat{T}(\mathbf{x})\|_\infty) \sum_\alpha \widehat{\mathcal{M}}(\alpha) \mathbb{1}_{\widehat{T}(\mathbf{x}) \in R_\alpha^\epsilon}.$$

Before investigating how the algorithm above empirically performs when applied to synthetic/real datasets, a few remarks are in order.

Remark 8.15. (INTERPRETATION OF THE PARAMETERS) In view of (8.39), 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 8.2) is already sparse (the number of charges cones is significantly less than 2^d).

Remark 8.16. (CHOICE OF PARAMETERS) A standard choice of parameters $(\epsilon,\ k,\ p)$ 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 8.12 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 novelty-detection 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 et al. (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 (8.1) and in view of Lemma 8.1.

Remark 8.17. (DIMENSION REDUCTION) If the extreme dependence structure is low dimensional, namely concentrated on low dimensional cones $\mathcal{C}_\alpha$ – or in other terms if only a limited number of margins can be large together – then most of the $\widehat{V}_i$'s will be concentrated on the R^ϵ_α's such that $|\alpha|$ (the dimension of the cone $\mathcal{C}_\alpha$) is small; then the representation of the dependence structure in (8.42) is both sparse and low dimensional.

Remark 8.18. (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.

8.5 Experimental results

8.5.1 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 $\{\mathcal{C}_{\alpha_m}, 1 \le m \le K\}$. 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 4. 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 8.1 shows the averaged numbers of errors among the 100 experiments. The results are very promising in situations where the number of sub-cones is moderate *w.r.t.* the number of observations.

TABLE 8.1: Support recovering on simulated data

# 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

8.5.2 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 8.2 show that the number of sub-cones $\mathcal{C}_\alpha$ identified by Algorithm 4 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 4 are given in Fig. 8.5.

fig_source/wave_dir2.png

FIGURE 8.5: sub-cone dimensions of wave data

8.5.3 Application to Anomaly Detection on real-world data sets

The main purpose of Algorithm 4 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)).

TABLE 8.2: Total number of sub-cones 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

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 novelty detection 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%.

The last three datasets belong to the KDD Cup '99 dataset (KDDCup (1999), Tavallaee 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 preprocessing 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 8.3 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 8.4. DAMEX significantly improves the performance (both in term of precision and of ROC curves) in extreme regions for each dataset, as illustrated in figures 8.6 and 8.7.

In Table 8.5, we repeat the same experiments but with $\epsilon = 0.1$. This yields the same strong performance of DAMEX, excepting for *SF* (see Figure 8.8). Generally, to large ϵ may yield

[1] These datasets are available for instance on http://scikit-learn.org/dev/

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 were 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 sub-cones (under-estimation of the asymptotic mass), which yields better performances.

TABLE 8.3: Datasets characteristics

	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 8.4: Results on extreme regions with standard parameters $(k, \epsilon) = (n^{1/2}, 0.01)$

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 8.5: Results on extreme regions with lower $\epsilon = 0.1$

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**

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 4 to treat new observations that appear fig_source/shuttle-semi-supervised-average-rect-01.png

FIGURE 8.6: shuttle, default parameters

fig_source/SF-4d-lb-semi-supervised-average-rect-01.png

FIGURE 8.7: SF dataset, default parameters

in the extreme region, and the standard algorithm to deal with those which appear in the non-extreme region.

8.6 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 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 fig_source/SF-4d-lb-semi-supervised-average-rect-1.png

FIGURE 8.8: SF dataset, larger ϵ

fig_source/SA-lb-semi-supervised-average-rect-01.png

FIGURE 8.9: SA dataset, default parameters

challenges in multivariate EVT. Since the asymptotic bias ($\mathrm{bias}(\alpha, n, k, \epsilon)$ in eq. (8.37)) 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 in an adaptive way with respect to k and n (see Remark 8.16). This might also open up the possibility of de-biasing the estimation procedure (Fougeres et al. (2015), Beirlant et al. (2016)). As a second contribution, this work extends the applicability of 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/forestcover-semi-supervised-average-rect-01.png

FIGURE 8.10: forestcover dataset, default parameters

fig_source/http-3d-semi-supervised-average-rect-01.png

FIGURE 8.11: http dataset, default parameters

8.7 Technical proofs

8.7.1 Proof of Lemma 8.5

For n vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ in $\mathbb{R}^d$, let us denote by $rank(v_i^j)$ the rank of v_i^j among $v_1^j, \dots, v_n^j$, that is $rank(v_i^j) = \sum_{k=1}^n \mathbb{1}_{\{v_k^j \leq v_i^j\}}$, so that $\hat{F}_j(X_i^j) = (rank(X_i^j) - 1)/n$. For the first equivalence, notice that $\hat{V}_i^j = 1/\hat{U}_i^j$. For the others, we have both at the same time:

$$\begin{aligned}
\hat{V}_i^j \geq \frac{n}{k} x_j \;&\Leftrightarrow\; 1 - \frac{rank(X_i^j) - 1}{n} \;\leq\; \frac{k}{n}\, x_j^{-1} \\
&\Leftrightarrow\; rank(X_i^j) \geq n - k x_j^{-1} + 1 \\
&\Leftrightarrow\; rank(X_i^j) \geq n - \lfloor k x_j^{-1} \rfloor + 1 \\
&\Leftrightarrow\; X_i^j \geq X^j_{(n - \lfloor k x_j^{-1} \rfloor + 1)},
\end{aligned}$$

and

$$\begin{aligned}
X_i^j \geq X^j_{(n - \lfloor k x_j^{-1} \rfloor + 1)} \;&\Leftrightarrow\; rank(X_i^j) \geq n - \lfloor k x_j^{-1} \rfloor + 1 \\
&\Leftrightarrow\; rank(F_j(X_i^j)) \geq n - \lfloor k x_j^{-1} \rfloor + 1 \qquad \text{(with probability one)} \\
&\Leftrightarrow\; rank(1 - F_j(X_i^j)) \leq \lfloor k x_j^{-1} \rfloor \\
&\Leftrightarrow\; U_i^j \leq U^j_{(\lfloor k x_j^{-1} \rfloor)}.
\end{aligned}$$

8.7.2 Proof of Lemma 8.6

First, recall that $g_{\alpha,\beta}(\mathbf{x}, \mathbf{z}) = \mu\big(R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha, \beta)\big)$, see (8.29). Denote by π the transformation to pseudo-polar coordinates introduced in Section 8.2,

$$\begin{aligned}
\pi : [0, \infty]^d \setminus \{\mathbf{0}\} &\to (0, \infty] \times S_\infty^{d-1} \\
\mathbf{v} &\mapsto (r, \boldsymbol{\theta}) = (\|\mathbf{v}\|_\infty, \|\mathbf{v}\|_\infty^{-1} \mathbf{v}).
\end{aligned}$$

Then, we have $\mathrm{d}(\mu \circ \pi^{-1}) = \frac{\mathrm{d}r}{r^2} \mathrm{d}\Phi$ 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 (8.5). Then

$$\begin{aligned}
g_{\alpha,\beta}(\mathbf{x}, \mathbf{z}) &= \mu \circ \pi^{-1}\Big\{(r, \boldsymbol{\theta}) : \quad \forall i \in \alpha,\ r\theta_i \geq x_i^{-1} \ ; \quad \forall j \in \beta, r\theta_j < z_j^{-1}\Big\} \\
&= \mu \circ \pi^{-1}\Big\{(r, \boldsymbol{\theta}) : \quad r \geq \bigvee_{i \in \alpha} (\theta_i x_i)^{-1} \ ; \quad r < \bigwedge_{j \in \beta} (\theta_j z_j)^{-1}\Big\} \\
&= \int_{\boldsymbol{\theta} \in S_\infty^{d-1}} \int_{r > 0} \mathbb{1}_{r \geq \bigvee_{i \in \alpha} (\theta_i x_i)^{-1}}\, \mathbb{1}_{r < \bigwedge_{j \in \beta} (\theta_j z_j)^{-1}} \frac{\mathrm{d}r}{r^2} \mathrm{d}\Phi(\boldsymbol{\theta}) \\
&= \int_{\boldsymbol{\theta} \in S_\infty^{d-1}} \left(\Big(\bigvee_{i \in \alpha} (\theta_i x_i)^{-1} \Big)^{-1} - \Big(\bigwedge_{j \in \beta} (\theta_j z_j)^{-1} \Big)^{-1} \right)_+ \mathrm{d}\Phi(\boldsymbol{\theta}) \\
&= \int_{\boldsymbol{\theta} \in S_\infty^{d-1}} \left(\bigwedge_{i \in \alpha} \theta_i x_i - \bigvee_{j \in \beta} \theta_j z_j \right)_+ \mathrm{d}\Phi(\boldsymbol{\theta}),
\end{aligned}$$

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 $\theta \in S_\infty^{d-1}$:

$$\begin{aligned}
\Big(\bigwedge_{j\in\alpha} \theta_j x_j - \bigvee_{j\in\beta} \theta_j z_j\Big)_+ - \Big(\bigwedge_{j\in\alpha} \theta_j x'_j - \bigvee_{j\in\beta} \theta_j z'_j\Big)_+ & \\
\leq \Big[\Big(\bigwedge_{j\in\alpha} \theta_j x_j - \bigvee_{j\in\beta} \theta_j z_j\Big) - \Big(\bigwedge_{j\in\alpha} \theta_j x'_j - \bigvee_{j\in\beta} \theta_j z'_j\Big)\Big]_+ & \\
\leq \Big[\bigwedge_{j\in\alpha} \theta_j x_j - \bigwedge_{j\in\alpha} \theta_j x'_j \;+\; \bigvee_{j\in\beta} \theta_j z'_j - \bigvee_{j\in\beta} \theta_j z_j\Big]_+ & \\
\leq \Big[\max_{j\in\alpha}(\theta_j x_j - \theta_j x'_j) \;+\; \max_{j\in\beta}(\theta_j z'_j - \theta_j z_j)\Big]_+ & \\
\leq \max_{j\in\alpha} \theta_j |x_j - x'_j| \;+\; \max_{j\in\beta} \theta_j |z'_j - z_j| &
\end{aligned}$$

Hence,

$$\begin{aligned}
&|g_{\alpha,\beta}(\mathbf{x},\mathbf{z}) - g_{\alpha,\beta}(\mathbf{x}',\mathbf{z}')| \\
&\qquad\leq \int_{S_\infty^{d-1}} \Big(\max_{j\in\alpha} \theta_j |x_j - x'_j| \;+\; \max_{j\in\beta} \theta_j |z'_j - z_j|\Big)\, \mathrm{d}\Phi(\boldsymbol{\theta})\,.
\end{aligned}$$

Now, by (8.6) we have:

$$\int_{S_\infty^{d-1}} \max_{j\in\alpha} \theta_j |x_j - x'_j| \;\, \mathrm{d}\Phi(\boldsymbol{\theta}) = \mu([\mathbf{0}, \tilde{\mathbf{x}}^{-1}]^c)$$

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, $\int_{S_\infty^{d-1}} \max_{j\in\beta} \theta_j |z'_j - z_j| \;\, \mathrm{d}\Phi(\boldsymbol{\theta}) \;\leq\; \sum_{j\in\beta} |z_j - z'_j|$.

8.7.3 Proof of Proposition 8.8

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 $\mathcal{C}_n(\,\cdot\,) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\mathbf{Z}_i \in\,\cdot\,\}}$ and $\mathcal{C}(\mathbf{x}) = \mathbb{P}(\mathbf{Z} \in\,\cdot\,)$ be the empirical and true measures associated with a n-sample $\mathbf{Z}_1, \ldots, \mathbf{Z}_d$ of $i.i.d.$ realizations of a random vector $\mathbf{Z} = (Z^1, \ldots, Z^d)$ with uniform margins on $[0,1]$. Then for any real number $\delta \geq e^{-k}$, with probability greater than $1 - \delta$,

$$\sup_{0\leq\mathbf{x}\leq T} \frac{n}{k} \left|\mathcal{C}_n(\frac{k}{n}[\mathbf{x},\infty[^c) - \mathcal{C}(\frac{k}{n}[\mathbf{x},\infty[^c)\right| \;\leq\; Cd\sqrt{\frac{T}{k}\log\frac{1}{\delta}}\,. \tag{8.43}$$

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 8.8 follows the same lines as in Goix et al. (2015b). The cornerstone concentration inequality (8.43) has to be replaced with

$$\max_{\alpha,\beta} \sup_{\substack{0 \le \mathbf{x},\mathbf{z} \le T \\ \exists j \in \alpha, x_j \le T'}} \frac{n}{k} \left| \mathcal{C}_n\left(\frac{k}{n} R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha,\ \beta)^{-1}\right) - \mathcal{C}\left(\frac{k}{n} R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha,\ \beta)^{-1}\right) \right|$$

$$\le Cd\sqrt{\frac{dT'}{k}\log\frac{1}{\delta}}\,. \tag{8.44}$$

Remark 8.19. Inequality (8.44) 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 (8.44) with $T = T'$ in the analysis below, since the smaller bounds in T' obtained (on $\Lambda(n)$ in (8.47)) would be diluted (by $\Upsilon(n)$ in (8.47)).

Proof of (8.44). 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

$$\mathcal{A} \;=\; \mathcal{A}_{T,T'} \;=\; \bigcup_{\alpha,\beta} \mathcal{A}_{T,T',\alpha,\beta} \quad \text{with}$$

$$\mathcal{A}_{T,T',\alpha,\beta} \;=\; \frac{k}{n}\Big\{ R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha,\ \beta)^{-1} :\ \mathbf{x}, \mathbf{z} \in \mathbb{R}^d,\ 0 \le \mathbf{x}, \mathbf{z} \le T,$$
$$\exists j \in \alpha, x_j \le T' \Big\}\,,$$

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 (8.26) that

$$R(\mathbf{x}^{-1}, \mathbf{z}^{-1}, \alpha,\ \beta)^{-1} \;=\; \Big\{ \mathbf{y} \in [0, \infty]^d,\ \ y_j \le x_j \ \ \text{for } j \in \alpha,$$
$$y_j > z_j \ \ \text{for } j \in \beta \ \Big\}$$
$$=\; [\mathbf{a}, \mathbf{b}],$$

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 $\mathbb{P}(\mathbf{Z} \in \mathbb{A}) \le \mathbb{P}(\mathbf{Z} \in [\frac{k}{n}\mathbf{T}',\ \infty[^c) \le \sum_{j=1}^d \mathbb{P}(Z^j \le \frac{k}{n}T') \le \frac{k}{n} dT'$. Inequality (8.44) is thus a direct consequence of Theorem 1 in Goix et al. (2015b). □

Define now the empirical version $\tilde{F}_{n,\alpha,\beta}$ of $\tilde{F}_{\alpha,\beta}$ (introduced in (8.28)) as

$$\tilde{F}_{n,\alpha,\beta}(\mathbf{x}, \mathbf{z}) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i^j \le x_j \ \text{ for } j \in \alpha \ \text{ and } \ U_i^j > z_j \ \text{ for } j \in \beta\}}\,, \tag{8.45}$$

so that $\frac{n}{k}\tilde{F}_{n,\alpha,\beta}(\frac{k}{n}\mathbf{x}, \frac{k}{n}\mathbf{z}) \;=\; \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{U_i^j \le \frac{k}{n}x_j \ \text{ for } j \in \alpha \ \text{ and } \ U_i^j > \frac{k}{n}z_j \ \text{ for } j \in \beta\}}$. 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 (8.30)) allowing to handle uniform variables. This is illustrated by the following lemmas.

Lemma 8.20 (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

$$g_{n,\alpha,\beta}(\mathbf{x},\mathbf{z}) \;=\; \frac{n}{k}\tilde{F}_{n,\alpha,\beta}\left(\left(U^j_{(\lfloor kx_j\rfloor)}\right)_{j\in\alpha},\left(U^j_{(\lfloor kz_j\rfloor)}\right)_{j\in\beta}\right),$$

Proof. Considering the definition in (8.45) and (8.31), both sides are equal to $\mu_n(R(\mathbf{x}^{-1},\mathbf{z}^{-1},\alpha,\beta))$. □

Lemma 8.21 (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:*

$$\max_{\alpha,\beta}\sup_{0\leq\mathbf{x},\mathbf{z}\leq T}\left|\frac{n}{k}\tilde{F}_{n,\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z})-\frac{n}{k}\tilde{F}_{\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z})\right|\leq Cd\sqrt{\frac{T}{k}\log\frac{1}{\delta}}\,.$$

Proof. Notice that

$$\begin{aligned}&\sup_{0\leq\mathbf{x},\mathbf{z}\leq T}\left|\frac{n}{k}\tilde{F}_{n,\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z})-\frac{n}{k}\tilde{F}_{\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z})\right|\\&=\sup_{0\leq\mathbf{x},\mathbf{z}\leq T}\frac{n}{k}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{U}_i\in\frac{k}{n}R(\mathbf{x}^{-1},\mathbf{z}^{-1},\alpha,\,\beta)^{-1}}-\mathbb{P}\left[\mathbf{U}\in\frac{k}{n}R(\mathbf{x}^{-1},\mathbf{z}^{-1},\alpha,\,\beta)^{-1}\right]\right|,\end{aligned}$$

and apply inequality (8.44) with $T'=T$. □

Remark 8.22. Note that the following stronger inequality holds true, when using (8.44) in full generality, *i.e.* with $T'<T$. For any finite $T,T'>0$, and $\delta\geq e^{-k}$, with probability at least $1-\delta$,

$$\max_{\alpha,\beta}\sup_{\substack{0\leq\mathbf{x},\mathbf{z}\leq T\\ \exists j\in\alpha,x_j\leq T'}}\left|\frac{n}{k}\tilde{F}_{n,\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z})-\frac{n}{k}\tilde{F}_{\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z})\right|\leq Cd\sqrt{\frac{T'}{k}\log\frac{1}{\delta}}.$$

The following lemma is stated and proved in Goix et al. (2015b).

Lemma 8.23 (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 $1-\delta$,*

$$\forall\, 1\leq j\leq d,\quad \frac{n}{k}U^j_{(\lfloor kT\rfloor)}\;\leq\;2T\,,\tag{8.46}$$

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

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

We may now proceed with the proof of Proposition 8.8. Using Lemma 8.20, we may write:

$$\begin{aligned}
&\max_{\alpha,\beta} \sup_{0\leq \mathbf{x},\mathbf{z}\leq T} |g_{n,\alpha,\beta}(\mathbf{x},\mathbf{z}) - g_{\alpha,\beta}(\mathbf{x},\mathbf{z})| \\
&\quad = \max_{\alpha,\beta} \sup_{0\leq \mathbf{x},\mathbf{z}\leq T} \left| \frac{n}{k}\tilde{F}_{n,\alpha,\beta}\left(\left(U^j_{(\lfloor kx_j\rfloor)}\right)_{j\in\alpha}, \left(U^j_{(\lfloor kz_j\rfloor)}\right)_{j\in\beta} \right) - g_{\alpha,\beta}(\mathbf{x},\mathbf{z}) \right| \\
&\quad \leq \Lambda(n) + \Xi(n) + \Upsilon(n) . \qquad (8.47)
\end{aligned}$$

with:

$$\begin{aligned}
\Lambda(n) &= \max_{\alpha,\beta} \sup_{0\leq \mathbf{x},\mathbf{z}\leq T} \Bigg| \frac{n}{k}\tilde{F}_{n,\alpha,\beta}\left(\left(U^j_{(\lfloor kx_j\rfloor)}\right)_{j\in\alpha}, \left(U^j_{(\lfloor kz_j\rfloor)}\right)_{j\in\beta} \right) \\
&\qquad\qquad - \frac{n}{k}\tilde{F}_{\alpha,\beta}\left(\left(U^j_{(\lfloor kx_j\rfloor)}\right)_{j\in\alpha}, \left(U^j_{(\lfloor kz_j\rfloor)}\right)_{j\in\beta} \right) \Bigg| \\
\Xi(n) &= \max_{\alpha,\beta} \sup_{0\leq \mathbf{x},\mathbf{z}\leq T} \Bigg| \frac{n}{k}\tilde{F}_{\alpha,\beta}\left(\left(U^j_{(\lfloor kx_j\rfloor)}\right)_{j\in\alpha}, \left(U^j_{(\lfloor kz_j\rfloor)}\right)_{j\in\beta} \right) \\
&\qquad\qquad - g_{\alpha,\beta}\left(\left(\frac{n}{k}U^j_{(\lfloor kx_j\rfloor)}\right)_{j\in\alpha}, \left(\frac{n}{k}U^j_{(\lfloor kz_j\rfloor)}\right)_{j\in\beta} \right) \Bigg| \\
\Upsilon(n) &= \max_{\alpha,\beta} \sup_{0\leq \mathbf{x},\mathbf{z}\leq T} \left| g_{\alpha,\beta}\left(\left(\frac{n}{k}U^j_{(\lfloor kx_j\rfloor)}\right)_{j\in\alpha}, \left(\frac{n}{k}U^j_{(\lfloor kz_j\rfloor)}\right)_{j\in\beta} \right) - g_{\alpha,\beta}(\mathbf{x},\mathbf{z}) \right| .
\end{aligned}$$

Now, considering (8.46) we have with probability greater than $1-\delta$ that for every $1\leq j\leq d$, $U^j_{(\lfloor kT\rfloor)} \leq 2T\frac{k}{n}$, 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 8.21, 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:

$$\begin{aligned}
\Upsilon(n) &\leq \max_{\alpha,\beta} \sup_{0\leq \mathbf{x},\mathbf{z}\leq T} \Bigg| g_{\alpha,\beta}\left(\frac{n}{k}\left(U^j_{(\lfloor kx_j\rfloor)}\right)_{j\in\alpha}, \frac{n}{k}\left(U^j_{(\lfloor kz_j\rfloor)}\right)_{j\in\beta} \right) \\
&\qquad\qquad - g_{\alpha,\beta}\left(\left(\frac{\lfloor kx_j\rfloor}{k}\right)_{j\in\alpha}, \left(\frac{\lfloor kz_j\rfloor}{k}\right)_{j\in\beta} \right) \Bigg| \\
&\quad + \max_{\alpha,\beta} \sup_{0\leq \mathbf{x},\mathbf{z}\leq T} \left| g_{\alpha,\beta}\left(\left(\frac{\lfloor kx_j\rfloor}{k}\right)_{j\in\alpha}, \left(\frac{\lfloor kz_j\rfloor}{k}\right)_{j\in\beta} \right) - g_{\alpha,\beta}(\mathbf{x},\mathbf{z}) \right| \\
&=: \Upsilon_1(n) + \Upsilon_2(n) .
\end{aligned}$$

The inequality in Lemma 8.6 allows us to bound the first term $\Upsilon_1(n)$:

$$\begin{aligned}
\Upsilon_1(n) &\leq C\max_{\alpha,\beta} \sup_{0\leq \mathbf{x},\mathbf{z}\leq T} \sum_{j\in\alpha} \left| \frac{\lfloor kx_j\rfloor}{k} - \frac{n}{k}U^j_{(\lfloor kx_j\rfloor)} \right| + \sum_{j\in\beta} \left| \frac{\lfloor kz_j\rfloor}{k} - \frac{n}{k}U^j_{(\lfloor kz_j\rfloor)} \right| \\
&\leq 2C \sup_{0\leq \mathbf{x}\leq T} \sum_{1\leq j\leq d} \left| \frac{\lfloor kx_j\rfloor}{k} - \frac{n}{k}U^j_{(\lfloor kx_j\rfloor)} \right|
\end{aligned}$$

so that by Lemma 8.23, with probability greater than $1-(d+1)\delta$:

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

Similarly,

$$\Upsilon_2(n) \;\le\; 2C \sup_{0\le\mathbf{x}\le T} \sum_{1\le j\le d}\left|\frac{\lfloor kx_j\rfloor}{k}-x_j\right| \;\le\; C\frac{2d}{k}\,.$$

Finally we get, for every $n>0$, with probability at least $1-(d+3)\delta$,

$$\begin{aligned}
\max_{\alpha,\beta}\sup_{0\le\mathbf{x},\mathbf{z}\le T}|g_{n,\alpha,\beta}(\mathbf{x},\mathbf{z})-g_{\alpha,\beta}(\mathbf{x},\mathbf{z})| \;&\le\; \Lambda(n)+\Upsilon_1(n)+\Upsilon_2(n)+\Xi(n)\\
&\le\; Cd\sqrt{\frac{2T}{k}\log\frac{1}{\delta}} \;+\; \frac{2d}{k} \;+\; \max_{\alpha,\beta}\sup_{0\le\mathbf{x},\mathbf{z}\le 2T}\left|\frac{n}{k}\tilde F_{\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z})-g_{\alpha,\beta}(\mathbf{x},\mathbf{z})\right|\\
&\le\; C'd\sqrt{\frac{2T}{k}\log\frac{1}{\delta}} \;+\; \max_{\alpha,\beta}\sup_{0\le\mathbf{x},\mathbf{z}\le 2T}\left|\frac{n}{k}\tilde F_{\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z})-g_{\alpha,\beta}(\mathbf{x},\mathbf{z})\right|.
\end{aligned}$$

Remark 8.24. (BIAS TERM) It is classical (see Qi (1997) p.174 for details) to extend the simple convergence (8.27) 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,

$$\sup_{0\le\mathbf{x},\mathbf{z}\le 2T}\left|\frac{n}{k}\tilde F_{\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z})-g_{\alpha,\beta}(\mathbf{x},\mathbf{z})\right|\to 0$$

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:

$$\max_{\alpha,\beta}\sup_{0\le\mathbf{x},\mathbf{z}\le 2T}\left|\frac{n}{k}\tilde F_{\alpha,\beta}(\frac{k}{n}\mathbf{x},\frac{k}{n}\mathbf{z})-g_{\alpha,\beta}(\mathbf{x},\mathbf{z})\right|\to 0. \tag{8.48}$$

8.7.4 Proof of Lemma 8.10

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 8.3), Φ is concentrated on the (disjoint) edges

$$\begin{aligned}
\Omega_{\alpha,i_0}=\{\mathbf{x}:\ \|\mathbf{x}\|_\infty=1,\ x_{i_0}=1,\ \ 0<x_i<1\ \ &\text{for } i\in\alpha\setminus\{i_0\}\\
x_i=0\qquad &\text{for } i\notin\alpha\quad\}
\end{aligned}$$

and that the restriction Φ_{α,i_0} of Φ to Ω_{α,i_0} is absolutely continuous *w.r.t.* the Lebesgue measure $\mathrm{d}x_{\alpha\setminus i_0}$ on the cube's edges, whenever $|\alpha|\ge 2$. By (8.15) we have, for every $\beta\supset\alpha$,

$$\begin{aligned}
\Phi(\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_\beta) \;&=\; \sum_{i_0\in\beta}\int_{\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_{\beta,i_0}}\frac{\mathrm{d}\Phi_{\beta,i_0}}{\mathrm{d}x_{\beta\setminus i_0}}(x)\,\mathrm{d}x_{\beta\setminus i_0}\\
\Phi(\Omega_\alpha) \;&=\; \sum_{i_0\in\alpha}\int_{\Omega_{\alpha,i_0}}\frac{\mathrm{d}\Phi_{\alpha,i_0}}{\mathrm{d}x_{\alpha\setminus i_0}}(x)\,\mathrm{d}x_{\alpha\setminus i_0}\,.
\end{aligned}$$

Thus,

$$\begin{aligned}\Phi(\Omega_\alpha^{\epsilon,\epsilon'}) - \Phi(\Omega_\alpha) &= \sum_{\beta\supset\alpha}\sum_{i_0\in\beta}\int_{\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_{\beta,i_0}} \frac{\mathrm{d}\Phi_{\beta,i_0}}{\mathrm{d}x_{\beta\setminus i_0}}(x)\,\mathrm{d}x_{\beta\setminus i_0}\\ &\qquad - \sum_{i_0\in\alpha}\int_{\Omega_{\alpha,i_0}} \frac{\mathrm{d}\Phi_{\alpha,i_0}}{\mathrm{d}x_{\alpha\setminus i_0}}(x)\,\mathrm{d}x_{\alpha\setminus i_0}\\ &= \sum_{\beta\supsetneq\alpha}\sum_{i_0\in\beta}\int_{\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_{\beta,i_0}} \frac{\mathrm{d}\Phi_{\beta,i_0}}{\mathrm{d}x_{\beta\setminus i_0}}(x)\,\mathrm{d}x_{\beta\setminus i_0}\\ &\qquad - \sum_{i_0\in\alpha}\int_{\Omega_{\alpha,i_0}\setminus(\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_{\alpha,i_0})} \frac{\mathrm{d}\Phi_{\alpha,i_0}}{\mathrm{d}x_{\alpha\setminus i_0}}(x)\,\mathrm{d}x_{\alpha\setminus i_0},\end{aligned}$$

so that by (8.16),

$$\begin{aligned}|\Phi(\Omega_\alpha^{\epsilon,\epsilon'}) - \Phi(\Omega_\alpha)| &\le \sum_{\beta\supsetneq\alpha} M_\beta \sum_{i_0\in\beta}\int_{\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_{\beta,i_0}} \mathrm{d}x_{\beta\setminus i_0}\\ &\qquad + M_\alpha\sum_{i_0\in\alpha}\int_{\Omega_{\alpha,i_0}\setminus(\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_{\alpha,i_0})} \mathrm{d}x_{\alpha\setminus i_0}\,.\end{aligned} \tag{8.49}$$

Without loss of generality we may assume that $\alpha = \{1, ..., K\}$ with $K \le d$. Then, for $\beta \supsetneq \alpha$, $\int_{\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_{\beta,i_0}} \mathrm{d}x_{\beta\setminus i_0}$ 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 $\beta = \{1, ..., P\}$ with $P > K$. Then

$$\begin{aligned}\Omega_\alpha^{\epsilon,\epsilon'}\ \cap\ \Omega_{\beta,i_0} = \{\epsilon < x_1, ..., x_K \le 1,\ & x_{K+1}, ..., x_P \le \epsilon',\ x_{i_0} = 1,\\ & x_{P+1} = ... = x_d = 0 \qquad \}\end{aligned}$$

which is empty if $i_0 \ge K+1$ (*i.e.* $i_0 \in \beta\setminus\alpha$) and which fulfills if $i_0 \le K$

$$\int_{\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_{\beta,i_0}} \mathrm{d}x_{\beta\setminus i_0} \le (\epsilon')^{P-K}.$$

The first term in (8.49) is then bounded by $\sum_{\beta\supsetneq\alpha} M_\beta|\alpha|(\epsilon')^{|\beta|-|\alpha|}$. Now, concerning the second term in (8.49), $\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_{\alpha,i_0} = \{\epsilon < x_1, ..., x_K \le 1, x_{i_0} = 1,\ x_{K+1}, ..., x_d = 0\}$ and then

$$\Omega_{\alpha,i_0}\setminus(\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_{\alpha,i_0}) = \bigcup_{l=1,\dots,K}\Omega_{\alpha,i_0}\cap\{x_l \le \epsilon\},$$

so that $\int_{\Omega_{\alpha,i_0}\setminus(\Omega_\alpha^{\epsilon,\epsilon'}\cap\Omega_{\alpha,i_0})} \mathrm{d}x_{\alpha\setminus i_0} \le K\epsilon = |\alpha|\epsilon$. The second term in (8.49) is thus bounded by $M|\alpha|^2\epsilon$. Finally, (8.49) implies

$$|\Phi(\Omega_\alpha^{\epsilon,\epsilon'}) - \Phi(\Omega_\alpha)| \le |\alpha|\sum_{\beta\supsetneq\alpha} M_\beta(\epsilon')^{|\beta|-|\alpha|} + M|\alpha|^2\epsilon.$$

To conclude, observe that by Assumption 3,

$$\sum_{\beta\supsetneq\alpha} M_\beta(\epsilon')^{|\beta|-|\alpha|} \le \sum_{\beta\supsetneq\alpha} M_\beta(\epsilon') \le \epsilon'\sum_{|\beta|\ge 2} M_\beta \le \epsilon' M$$

The result is thus proved.

8.7.5 Proof of Remark 8.14

Let us prove that Z_n, conditionally to the event $\{\|\frac{k}{n}V_1\|_\infty \geq 1\}$, 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 $\Phi : \mathbb{R}^{2^d-1} \to \mathbb{R}_+$ 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 $\pi_n \to \mu([\mathbf{0},\mathbf{1}]^c)$, 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 $\frac{n}{k}\mathbb{P}\left[Z_n(\alpha) = 1\right] = \frac{n}{k}\mathbb{P}\left[\frac{k}{n}\mathbf{V}_1 \in R_\alpha^\epsilon\right] \to \mu(R_\alpha^\epsilon)$, we find that

$$\mathbb{E}\left[\Phi(Z_n) \mid \|\frac{k}{n}\mathbf{V}_1\|_\infty \geq 1\right] \to \sum_\alpha \Phi(1_\alpha)\frac{\mu(R_\alpha^\epsilon)}{\mu([\mathbf{0},\mathbf{1}]^c)},$$

which achieves the proof.

PART IV

Efficient heuristic approaches

CHAPTER 9
How to Evaluate the Quality of Anomaly Detection Algorithms?

We recall that this is a contribution of heuristic nature and not yet supported by statistically sound theoretical results. This ongoing work has not been published yet and will certainly be completed in the near future, but we believe that it has its place in our manuscript, given the convincing empirical experiments and the rationale behind the approach promoted we gave.

Abstract This chapter presents the details relative to the introducing section 1.5.1.
When sufficient labeled data are available, classical criteria based on *Receiver Operating Characteristic* (ROC) or *Precision-Recall* (PR) 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. In this work, two criteria that do not require labels are empirically shown to discriminate accurately (*w.r.t.* ROC or PR based criteria) between algorithms. These criteria are based on existing Excess-Mass (EM) and Mass-Volume (MV) curves, which generally cannot be well estimated in large dimension. A methodology based on feature sub-sampling and aggregating is also described and tested, extending the use of these criteria to high-dimensional datasets and solving major drawbacks inherent to standard EM and MV curves.

Note: The material of this chapter is based on previous work published in Goix (2016) and on the submitted work Goix & Thomas (2016).

9.1 Introduction

When labels are available, classical ways to evaluate the quality of an anomaly scoring function are the ROC and PR curves. Unfortunately, most of the time, data come without any label. In lots of industrial setups, labeling datasets calls for costly human expertise, while more and more unlabeled data are available. A huge practical challenge is therefore to have access to criteria able to discriminate between unsupervised algorithms without using any labels. In this chapter, we formalize and justify the use of two such criteria designed for unsupervised anomaly detection, and adapt them to large dimensional data. Strong empirical performance demonstrates the relevance of our approach.

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 that generating the vast majority of the data. The usual assumption (in supervised learning) stipulating that the dataset contains structural information regarding all classes breaks down (Roberts, 1999): the very small number of points representing the abnormal class does not allow to learn information about this class. 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. 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. Classical parametric techniques (Barnett & Lewis, 1994; Eskin, 2000) assume that the normal data are generated by a distribution belonging to some specific and *a priori* known parametric model. The most popular non-parametric approaches include algorithms based on density (level set) estimation (Schölkopf et al., 2001; Scott & Nowak, 2006; Breunig et al., 2000), on dimensionality reduction (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); Markou & Singh (2003) for overviews of current research on anomaly detection.

It turns out that the overwhelming majority of anomaly detection algorithms return more than a binary label, normal/abnormal. They first compute a *scoring function*, which is converted to a binary prediction, typically by imposing some threshold based on its statistical distribution.

What is a scoring function? From a probabilistic point of view, there are different ways 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 if a new observation $\mathbf{x}$ has been generated from F, or from G. The optimal way to resolve this problem would be the likelihood ratio test, also called Neyman-Pearson test. If $(\mathrm{d}F/\mathrm{d}G)(\mathbf{x}) > t$ with $t > 0$ some threshold, then $\mathbf{x}$ has been drawn from F. Otherwise, $\mathbf{x}$ has been drawn from G. As anomalies are very rare, their structure cannot be observed in the data, in particular their distribution G. It is common and convenient (Vert, 2006) to replace G in the problem above by the Lebesgue measure, so that it boils down to estimating density level sets of F. This setup is typically the one of the One-Class Support Vector Machine (OCSVM) algorithm developed in Schölkopf et al. (2001), which extends the SVM methodology (Cortes & Vapnik, 1995; Shawe-Taylor & Cristianini, 2004) to handle training using only positive information. The underlying assumption is that we observe data in $\mathbb{R}^d$ from the normal class only, with underlying distribution F and underlying density $f : \mathbb{R}^d \to \mathbb{R}$. The goal is to estimate density level sets $(\{\mathbf{x}, f(\mathbf{x}) > t\})_{t>0}$ with t close to 0. In practice, such estimates are represented by a *scoring function*: any measurable function $s : \mathbb{R}^d \to \mathbb{R}_+$ integrable *w.r.t.* the Lebesgue measure $\mathrm{Leb}(.)$, *whose level sets are estimates of the true density level sets*. Any scoring function defines a pre-order 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.

How to know if a scoring function is good? How can we know if the pre-order induced by a scoring function s is 'close' to that of f, or equivalently if these induced level sets are close to those of f? The problem is to define this notion of proximity into a criterion $\mathcal{C}$, optimal scoring functions s^* being then defined as those optimizing $\mathcal{C}$. It turns out that for any strictly increasing transform $T : \mathrm{Im(f)} \to \mathbb{R}$, the level sets of $T \circ f$ are exactly those of f. Here and hereafter, Im(f) denotes the image of the mapping f. For instance, $2f$ or f^2 are perfect scoring functions, just as f. Thus, we cannot simply consider a criterion based on the distance of s to the true density, *e.g.* $\mathcal{C}(s) = \|s - f\|$. We seek for a similar criterion which is invariant by increasing transformation of the output s. In other words, the criterion should be defined in such a way that the collection of level sets of an optimal scoring function $s^*(x)$ coincides with that related to f. Moreover, any increasing transform of the density should be optimal regarding $\mathcal{C}$.

In the literature, two functional criteria admissible with respect to these requirements have been introduced: the Mass-Volume (MV) (Clémençon & Jakubowicz, 2013; Clémençon & Robbiano, 2014) and the Excess-Mass (EM) (Goix et al., 2015c) curves. Formally, it allows to consider $\mathcal{C}^{\Phi}(s) = \|\Phi(s) - \Phi(f)\|$ (instead of $\|s - f\|$) with $\Phi : \mathbb{R} \to \mathbb{R}_+$ 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. While such 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.

In this work, we define two numerical performance criteria based on MV and EM curves, which are tested with respect to three classical anomaly detection algorithms. A wide range on real labeled datasets are used in the benchmark. In addition, we propose a method based on feature sub-sampling and aggregating. It allows to scale this methodology to high-dimensional data which we use on the higher-dimensional datasets. We compare the results to ROC and PR criteria, which use the data labels hidden to MV and EM curves.

This chapter is structured as follows. Section 9.2 introduces Excess-Mass and Mass-Volume curves and defines associated numerical criteria. In Section 9.3, the feature sub-sampling based methodology to extend their use to high dimension is described. Finally, experiments on a wide range of real datasets are provided in Section 9.4.

9.2 Mass-Volume and Excess-Mass based criteria

We place ourselves in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We observe n $i.i.d.$ realizations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of a random variable $\mathbf{X} : \Omega \to \mathbb{R}^d$ representing the normal behavior, with *c.d.f.* F and density f *w.r.t.* the Lebesgue measure on $\mathbb{R}^d$. We denote by $\mathcal{S}$ the set of all scoring functions, namely any measurable function $s : \mathbb{R}^d \to \mathbb{R}_+$ integrable *w.r.t.* the Lebesgue measure. We work under the assumptions that the density f has no flat parts and is bounded. Excess-Mass and Mass-Volume curves are here introduced in a different way they originally were in Clémençon & Jakubowicz (2013); Goix et al. (2015c). We use equivalent definitions for them since the original definitions were more adapted to the ERM paradigm than to the issues addressed here.

9.2.1 Preliminaries

Let $s \in \mathcal{S}$ be a scoring function. In this context (Clémençon & Jakubowicz, 2013; Goix et al., 2015c), the mass-volume (MV) and the excess-mass (EM) curves of s can be written as

$$\forall \alpha \in (0,1), \quad MV_s(\alpha) = \inf_{u \geq 0} \ \text{Leb}(s \geq u) \quad \textit{s.t.} \quad \mathbb{P}(s(\mathbf{X}) \geq u) \geq \alpha \tag{9.1}$$

$$\forall t > 0, \quad EM_s(t) = \sup_{u \geq 0} \ \{\mathbb{P}(s(\mathbf{X}) \geq u) \ - \ t\text{Leb}(s \geq u)\} \tag{9.2}$$

The optimal curves are $MV^* = MV_f = MV_{T\circ f}$ and $EM^* = EM_f = EM_{T\circ f}$ for any increasing transform T : Im(f) $\to \mathbb{R}$. It can be proven (Clémençon & Jakubowicz, 2013; Goix et al., 2015c) that for any scoring function s, $MV^*(\alpha) \leq MV_s(\alpha)$ for all $\alpha \in (0,1)$ and $EM^*(t) \geq EM_s(t)$ for all $t > 0$. Also, $MV^*(\alpha)$ is the optimal value of the constrained minimization problem

$$\min_{\Gamma \text{ borelian}} \text{Leb}(\Gamma) \quad \textit{s.t.} \quad \mathbb{P}(\mathbf{X} \in \Gamma) \geq \alpha. \tag{9.3}$$

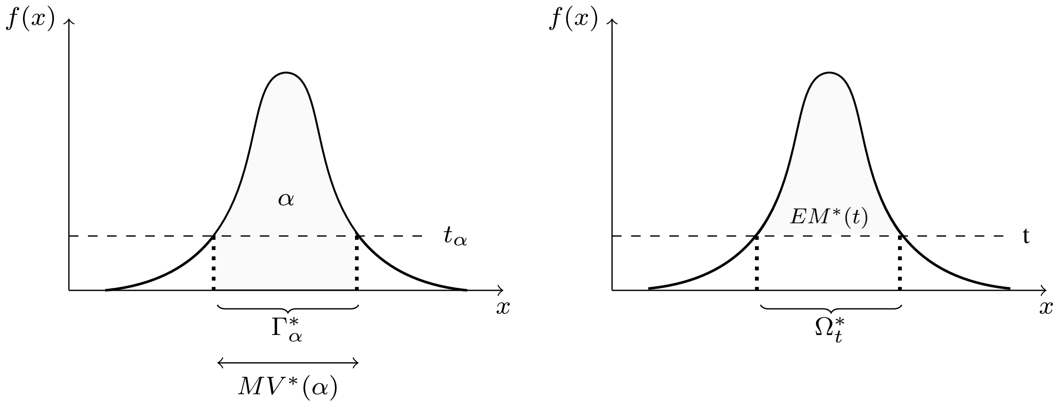
The minimization problem (9.3) has a unique solution Γ^*_α of mass α exactly, referred to as *minimum volume set* (Polonik, 1997): $MV^*(\alpha) = \text{Leb}(\Gamma^*_\alpha)$ and $\mathbb{P}(\mathbf{X} \in \Gamma^*_\alpha) = \alpha$.

Similarly, the optimal EM curve is linked with the notion of density excess-mass (as introduced in the seminal contribution Polonik (1995)). The main idea is to consider a Lagrangian formulation of the constrained minimization problem obtained by exchanging constraint and objective in (9.3),

$$EM^*(t) := \max_{\Omega \text{ borelian}} \{\mathbb{P}(\mathbf{X} \in \Omega) - t\text{Leb}(\Omega)\}. \tag{9.4}$$

Figure 9.1 compares the mass-volume and excess-mass approaches.

FIGURE 9.1: Comparison between $MV^*(\alpha)$ and $EM^*(t)$



9.2.2 Numerical unsupervised criteria

The main advantage of EM compared to MV is that the area under its curve is finite, even if the support of the distribution F is not. As curves cannot be trivially compared, consider the L^1-norm $\|.\|_{L^1(I)}$ with $I \subset \mathbb{R}$ an interval. As $MV^* = MV_f$ is below MV_s point-wise, $\arg\min_s \|MV_s - MV^*\|_{L^1(I)} = \arg\min \|MV_s\|_{L^1(I)}$. We thus define the criterion $\mathcal{C}^{MV}(s) = \|MV_s\|_{L^1(I^{MV})}$, 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 at one 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 considering $I^{EM} = [0, EM^{-1}(0.9)]$, with $EM_s^{-1}(0.9) := \inf\{t \geq 0,\ EM_s(t) \leq 0.9\}$, as $EM_s(0)$ is finite (equal to 1). We point out that such small values of t correspond to large level-sets. Also, we have observed that $EM_s^{-1}(0.9)$ (as well as $EM_f^{-1}(0.9)$) varies significantly depending on the dataset. Generally, for datasets in large dimension, it can be very small (in the experiments, smallest values are of order 10^{-7}) as it is of the same order of magnitude as the inverse of the total support volume.

As the distribution F of the normal data is generally unknown, mass-volume and excess-mass curves must be estimated. Let $s \in \mathcal{S}$ and $\mathbf{X}_1, \ \ldots, \ \mathbf{X}_n$ 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 (9.1) and (9.2),

$$\widehat{MV}_s(\alpha) = \inf_{u \geq 0} \text{Leb}(s \geq u) \quad \textit{s.t.} \quad \mathbb{P}_n(s \geq u) \geq \alpha \tag{9.5}$$

$$\widehat{EM}_s(t) = \sup_{u \geq 0} \mathbb{P}_n(s \geq u) \ - \ t\text{Leb}(s \geq u) \tag{9.6}$$

Note that in practice, the volume Leb$(s \geq u)$ is estimated using Monte-Carlo approximation, which only applies to small dimensions. Finally, we obtain the empirical EM and MV based performance criteria:

$$\widehat{\mathcal{C}}^{EM}(s) = \|\widehat{EM}_s\|_{L^1(I^{EM})} \qquad I^{EM} = [0, \widehat{EM}^{-1}(0.9)], \tag{9.7}$$

$$\widehat{\mathcal{C}}^{MV}(s) = \|\widehat{MV}_s\|_{L^1(I^{MV})} \qquad I^{MV} = [0.9, 0.999], \tag{9.8}$$

Remark 9.1. (LINK WITH ROC CURVE) To evaluate unsupervised algorithms, it is common to generate uniform outliers and then use the ROC curve approach. Up to identify the Lebesgue measure of a set to its empirical version (*i.e.* the proportion of uniform point inside), this approach is equivalent to using the mass-volume curve (Clémençon & Robbiano, 2014). However, in the former approach, the volume estimation does not appear directly, so that the (potentially huge) amount of uniform points needed to provide a good estimate of a volume is often not respected, yielding optimistic performances.

9.3 Scaling with dimension

In this section we propose a methodology to scale the use of the excess-mass and mass-volume criteria to large dimensional data. It 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 $F_1, \ldots, F_m$. A partial score $\widehat{\mathcal{C}}_k^{MV}$ (resp. $\widehat{\mathcal{C}}_k^{EM}$) is computed for each draw F_k using (9.7) (resp. (9.8)). The final performance criteria are obtained by averaging these partial criteria along the different draws of features. This methodology is described in Algorithm 5.

A drawback from this approach is that we do not evaluate combinations of more than d' features within the dependence structure. However, according to our experiments, this is enough in most of the cases. Besides, we solve two major drawbacks inherent to mass-volume or excess-mass criteria, which come from the Lebesgue reference measure:

- EM or MV performance criteria cannot be estimated in large dimension,
- EM or MV performance criteria cannot be compared when produced from spaces of different dimensions, since reference measures of $\mathbb{R}^d$ and $\mathbb{R}^{d+1}$ cannot be compared.

Algorithm 5 Evaluate anomaly detection algorithms on high dimensional data

```
Inputs: anomaly detection algorithm $\mathcal{A}$, data set $X = (x_i^j)_{1 \le i \le n, 1 \le j \le d}$, 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 $\widehat{s}_k = \mathcal{A}\big((x_i^j)_{1 \le i \le n,\ j \in F_k}\big)$
    compute $\widehat{\mathcal{C}}_k^{EM} = \|\widehat{EM}_{\widehat{s}_k}\|_{L^1(I^{EM})}$ using (9.7) or $\widehat{\mathcal{C}}_k^{MV} = \|\widehat{MV}_{\widehat{s}_k}\|_{L^1(I^{MV})}$ using (9.8)
end for

Return performance criteria:
```

$$\widehat{\mathcal{C}}^{EM}_{high\_dim}(\mathcal{A}) = \frac{1}{m}\sum_{k=1}^{m} \widehat{\mathcal{C}}_k^{EM} \quad \text{or} \quad \widehat{\mathcal{C}}^{MV}_{high\_dim}(\mathcal{A}) = \frac{1}{m}\sum_{k=1}^{m} \widehat{\mathcal{C}}_k^{MV} .$$

Remark 9.2. (FEATURE IMPORTANCE) With standard MV and EM curves, the benefit of using or not some feature j in training *cannot* be evaluated, as it involves spaces of different dimensions (d and $d+1$). Solving the second drawback precisely allows to evaluate the importance of features. By sub-sampling features, we can compare accuracies with or without using feature j: when computing $\widehat{\mathcal{C}}^{MV}_{high\_dim}$ or $\widehat{\mathcal{C}}^{EM}_{high\_dim}$ using Algorithm 5, this is reflected in the fact that j can (resp. cannot) be drawn.

Remark 9.3. (THEORETICAL GROUNDS) Criteria $\widehat{\mathcal{C}}^{MV}_{high\_dim}$ or $\widehat{\mathcal{C}}^{EM}_{high\_dim}$ do not evaluate a specific scoring function s produced by some algorithm (on some dataset), but the algorithm itself *w.r.t.* the dataset at stake. Indeed, these criteria proceed with the average of partial scoring functions on sub-space of $\mathbb{R}^d$. We have no theoretical guaranties that the final score does correspond to some scoring function defined on $\mathbb{R}^d$. In this work, we only show that from a practical point of view, it is a useful and accurate methodology to compare algorithms performance on large dimensional datasets.

Remark 9.4. (DEFAULT PARAMETERS) In our experiments, we arbitrarily chose $m = 50$ and $d' = 5$. This means that 50 draws of 5 features (with replacement after each draw) have been done. Volume in spaces of dimension 5 have thus to be estimated (which is feasible with Monte-Carlo), and 50 scoring functions (on random subspaces of dimension 5) have to be computed by the algorithm we want to evaluate. The next section shows (empirically) that these parameters achieve a good accuracy on the collection of datasets studied, the largest dimension considered being 164.

9.4 Benchmarks

Does performance in term of EM/MV correspond to performance in term of ROC/PR? Can we recover, on some fixed dataset, which algorithm is better than the others (according to ROC/PR criteria) without using labels? In this section we study four different empirical evaluations (ROC, PR, EM, MV) of three classical state-of-the-art anomaly detection algorithms, One-Class SVM (Schölkopf et al., 2001), Isolation Forest (Liu et al., 2008), and Local Outlier Factor (LOF) algorithm (Breunig et al., 2000), on 12 well-known anomaly detection datasets. Two criteria use labels (ROC and PR based criteria) and two do not (EM and MV based criteria). For ROC and PR curves, we consider the area under the (full) curve (AUC). For the excess-mass curve $EM(t)$ (resp. mass-volume curve), we consider the area under the curve on the interval $[0, EM^{-1}(0.9)]$ (resp. $[0.9, 0.999]$) as described in Section 9.2.

9.4.1 Datasets description

TABLE 9.1: Original Datasets characteristics

	nb of samples	nb of features	anomaly class	
adult	48842	6	class ’$> 50K$’	(23.9%)
http	567498	3	attack	(0.39%)
pima	768	8	pos (class 1)	(34.9%)
smtp	95156	3	attack	(0.03%)
wilt	4839	5	class ’w’ (diseased trees)	(5.39%)
annthyroid	7200	6	classes $\neq$ 3	(7.42%)
arrhythmia	452	164	classes $\neq$ 1 (features 10-14 removed)	(45.8%)
forestcover	286048	10	class 4 (vs. class 2)	(0.96%)
ionosphere	351	32	bad	(35.9%)
pendigits	10992	16	class 4	(10.4%)
shuttle	85849	9	classes $\neq$ 1 (class 4 removed)	(7.17%)
spambase	4601	57	spam	(39.4%)

The characteristics of these reference datasets are summarized in Table 9.1. They are all available on the UCI repository (Lichman, 2013) and the preprocessing is done in a classical way. We removed all non-continuous attributes as well as attributes taking less than 10 different values. The *http* and *smtp* datasets belong to the KDD Cup ’99 dataset (KDDCup, 1999; Tavallaee et al., 2009), which consists of a wide variety of hand-injected attacks (anomalies) in a closed network (normal background). They are classically obtained as described in Yamanishi et al. (2000). These datasets are available on the *scikit-learn* library (Pedregosa et al., 2011). The *shuttle* dataset is the fusion of the training and testing datasets available in the UCI repository. As in Liu et al. (2008), we use instances from all different classes but class 4. In the *forestcover* data, the normal data are the instances from class 2 while instances from class 4 are anomalies (as in Liu et al. (2008)). The *ionosphere* dataset differentiates ‘good’ from ‘bad’ radars, considered here as abnormal. A ‘good’ radar shows evidence of some type of structure in the ionosphere. A ‘bad’ radar does not, its signal passing through the ionosphere. The *spambase* dataset consists of spam or non-spam emails. The former constitute the abnormal class. The *annthyroid* medical dataset on hypothyroidism contains one normal class and two abnormal ones, which form the outlier set. The *arrhythmia* dataset reflects the presence and absence (class 1) of cardiac arrhythmia. The number of attributes being large considering the sample size, we removed attributes containing missing data. The *pendigits* dataset contains 10 classes corresponding to the digits from 0 to 9, examples being handwriting samples. As in Schubert et al. (2012), the abnormal data are chosen to be those from class 4. The *pima* dataset consists of medical data on diabetes. Patients suffering from diabetes (positive class) were considered outliers. The *wild* dataset involves detecting diseased trees in Quickbird imagery. Diseased trees (class ‘w’) is the abnormal class. In the *adult* dataset, the goal is to predict whether income exceeds $ 50K/year based on census data. Only the 6 continuous attributes are kept.

9.4.2 Results

The experiments are performed both in a novelty detection framework (also named semi-supervised framework, the training set consisting of normal data only) and in an unsupervised framework (the training set is polluted by abnormal data). In the former case, we simply removed anomalies from the training data, and EM and PR criteria are estimated using only normal data. In the latter case, the anomaly rate is arbitrarily bounded to 10% max, and EM and PR criteria are estimated with the same test data used for ROC and PR curves, without using their labels.

TABLE 9.2: Results for the novelty detection setting. ROC, PR, EM, MV often do agree on which algorithm is the best (in bold), which algorithm is the worse (underlined) on some fixed datasets. When they do not agree, it is often because ROC and PR themselves do not, meaning that the ranking is not clear.

Dataset	iForest				OCSVM				LOF			
	ROC	PR	EM	MV	ROC	PR	EM	MV	ROC	PR	EM	MV
adult	**0.661**	**0.277**	**1.0e-04**	**7.5e01**	0.642	0.206	2.9e-05	4.3e02	0.618	0.187	1.7e-05	9.0e02
http	0.994	0.192	1.3e-03	9.0	**0.999**	**0.970**	**6.0e-03**	**2.6**	0.946	0.035	8.0e-05	3.9e02
pima	0.727	0.182	5.0e-07	**1.2e04**	**0.760**	**0.229**	**5.2e-07**	1.3e04	0.705	0.155	3.2e-07	2.1e04
smtp	0.907	0.005	1.8e-04	9.4e01	0.852	**0.522**	**1.2e-03**	8.2	**0.922**	0.189	1.1e-03	**5.8**
wilt	0.491	0.045	4.7e-05	2.1e03	0.325	0.037	**5.9e-05**	**4.5e02**	**0.698**	**0.088**	2.1e-05	1.6e03
annthyroid	**0.913**	**0.456**	**2.0e-04**	2.6e02	0.699	0.237	6.3e-05	**2.2e02**	0.823	0.432	6.3e-05	1.5e03
arrhythmia	**0.763**	**0.487**	**1.6e-04**	**9.4e01**	0.736	0.449	1.1e-04	1.0e02	0.730	0.413	8.3e-05	1.6e02
forestcov.	0.863	0.046	3.9e-05	2.0e02	0.958	0.110	5.2e-05	1.2e02	**0.990**	**0.792**	**3.5e-04**	**3.9e01**
ionosphere	0.902	0.529	9.6e-05	7.5e01	**0.977**	**0.898**	**1.3e-04**	**5.4e01**	0.971	0.895	1.0e-04	7.0e01
pendigits	0.811	0.197	2.8e-04	2.6e01	0.606	0.112	2.7e-04	2.7e01	**0.983**	**0.829**	**4.6e-04**	**1.7e01**
shuttle	0.996	0.973	1.8e-05	5.7e03	0.992	0.924	**3.2e-05**	**2.0e01**	**0.999**	**0.994**	7.9e-06	2.0e06
spambase	**0.824**	**0.371**	**9.5e-04**	**4.5e01**	0.729	0.230	4.9e-04	1.1e03	0.754	0.173	2.2e-04	4.1e04

Recall that standard excess-mass and mass-volume performance criteria referring on the Lebesgue measure, they require volume estimation. They only apply to continuous datasets, of small dimension ($d \leq 8$). The datasets verifying these requirements are *http*, *smtp*, *pima*, *wilt* and *adult*. For the other datasets, we use the performance criteria $\widehat{C}^{MV}_{high\_dim}$ and $\widehat{C}^{EM}_{high\_dim}$ computed with Algorithm 5. We arbitrarily chose $m = 50$ and $d' = 5$, which means that 50 draws of 5 features, with replacement after each draw, are done. Other parameters have also been tested but are not presented here. The default parameters proposed here are a compromise between computational time and performance, in particular on the largest dimensional datasets. The latter require a relatively large product $m \times d'$, which is the maximal number of different features that can be drawn.

Excess-Mass, Mass-volume, ROC and Precision-Recall curves AUCs are presented in Table 9.2 for the novelty detection framework, and in Table 9.3 for the unsupervised framework. The corresponding ROC and PR curves are available at the end of this chapter. Figure 9.2 shows excess-mass and mass-volume curves on the adult dataset in a novelty detection setting. Corresponding figures for the other datasets are also available at the end of this chapter.

Results from Table 9.2 can be summarized as follows. Consider the 36 possible pairwise comparisons between the three algorithms over the twelve datasets

$$\left\{\left(A_1 \text{ on } \mathcal{D}, A_2 \text{ on } \mathcal{D}\right),\ A_1, A_2 \in \{\text{iForest, LOF, OCSVM}\},\ \mathcal{D} \in \{\text{adult, http}, \ldots, \text{spambase}\}\right\}. \tag{9.9}$$

TABLE 9.3: Results for the unsupervised setting still remain good: one can see that ROC, PR, EM, MV often do agree on which algorithm is the best (in bold), which algorithm is the worse (underlined) on some fixed datasets. When they do not agree, it is often because ROC and PR themselves do not, meaning that the ranking is not clear.

Dataset	iForest				OCSVM				LOF			
	ROC	PR	EM	MV	ROC	PR	EM	MV	ROC	PR	EM	MV
adult	**0.644**	**0.234**	**6.6e-05**	**2.7e02**	0.627	0.184	1.8e-05	5.6e02	0.545	0.098	7.4e-06	1.9e03
http	**0.999**	**0.686**	1.4e-03	2.2e01	0.994	0.207	**5.7e-03**	**3.3**	0.354	0.019	9.8e-05	3.9e02
pima	**0.747**	0.205	**1.2e-06**	**1.2e04**	0.742	**0.211**	6.0e-07	1.9e04	0.686	0.143	6.0e-07	3.2e04
smtp	0.902	0.004	2.7e-04	8.6e01	0.852	**0.365**	**1.4e-03**	7.7	**0.912**	0.057	1.1e-03	**7.0**
wilt	0.443	0.044	3.7e-05	2.2e03	0.318	0.036	**3.9e-05**	**4.3e02**	**0.620**	**0.066**	2.0e-05	8.9e02
annthyroid	**0.820**	**0.309**	**6.9e-05**	7.7e02	0.682	0.187	4.1e-05	**3.1e02**	0.724	0.175	1.6e-05	4.1e03
arrhythmia	**0.740**	0.416	**8.4e-05**	**1.1e02**	0.729	**0.447**	6.8e-05	1.2e02	0.729	0.409	5.6e-05	1.5e02
forestcov.	0.882	0.062	3.2e-05	2.3e02	**0.951**	**0.095**	4.4e-05	1.4e02	0.542	0.016	**2.4e-04**	**4.6e01**
ionosphere	0.895	0.543	7.4e-05	9.3e01	**0.977**	**0.903**	**8.7e-05**	**7.7e01**	0.969	0.884	6.9e-05	1.0e02
pendigits	0.463	0.077	2.7e-04	2.5e01	0.366	0.067	2.6e-04	2.8e01	**0.504**	**0.089**	**4.5e-04**	**1.6e01**
shuttle	**0.997**	**0.979**	7.1e-07	1.2e05	0.992	0.904	**5.8e-06**	**1.7e02**	0.526	0.116	7.1e-07	1.7e07
spambase	**0.799**	**0.303**	**2.2e-04**	**3.5e01**	0.714	0.214	1.5e-04	2.9e02	0.670	0.129	3.7e-05	2.7e04

FIGURE 9.2: MV and EM curves for adult dataset (novelty detection framework). Both in terms of EM and MV curves, iForest is found to perform better than OCSVM, which is itself found to perform better than LOF. Comparing to Table 9.2, ROC and PR AUCs give the same ranking (iForest on adult $\succ$ OCSVM on adult $\succ$ LOF on adult). The 3 pairwise comparisons (iForest on adult, LOF on adult), (OCSVM on adult, LOF on adult) and (OCSVM on adult, iForest on adult) are then similarly ordered by EM, PR, MV and EM criteria.

fig_source/evaluation_fig/mv_em_adult_supervised_09_factorized_inkscape.pn

For each dataset $\mathcal{D}$, there are three possible pairs (iForest on $\mathcal{D}$, LOF on $\mathcal{D}$), (OCSVM on $\mathcal{D}$, LOF on $\mathcal{D}$) and (OCSVM on $\mathcal{D}$, iForest on $\mathcal{D}$). Then the EM-score discriminates 28 of them (78%) as ROC score does, and 29 (81%) of them as PR score does. Intuitively this can be interpreted as follows. Choose randomly a dataset $\mathcal{D}$ among the twelve available, and two algorithms A_1, A_2 among the three available. This amounts to choose at random a pairwise comparison (A_1 on $\mathcal{D}$, A_2 on $\mathcal{D}$) among the 36 available. Suppose that according to ROC criterion, A_1 is better than A_2 on dataset $\mathcal{D}$, *i.e.* (A_1 on $\mathcal{D}$) $\succ$ (A_2 on $\mathcal{D}$). Then the EM-score discriminates A_1 and A_2 on dataset $\mathcal{D}$ in the same way, *i.e.* also finds A_1 to be better than A_2 on dataset $\mathcal{D}$ with 78 percent chance.

Besides, let us consider pairs (A_1 on $\mathcal{D}$, A_2 on $\mathcal{D}$) which are similarly ordered by ROC and PR criteria, namely *s.t.* A_1 is better than A_2 (or the reverse) on dataset $\mathcal{D}$ according to both EM and PR. According to Table 9.2, this represents every pairs but one in *spambase* and two in *smtp*. Then, one achieves $27/33 = 82\%$ of similarly discriminated pairs (*w.r.t.* to ROC and PR criteria). Moreover, EM is able to recover the exact (*w.r.t.* ROC and PR criteria) ranking of (A_1 on $\mathcal{D}$, A_2 on $\mathcal{D}$, A_3 on $\mathcal{D}$) on every datasets $\mathcal{D}$ excepting *wilt* and *shuttle*. For *shuttle*, note that ROC scores are very close to each other (0.996, 0.992, 0.999) and thus not clearly discriminates algorithms. The only significant error committed by EM is for the *wilt* dataset (on which no feature sub-sampling is done due to the low dimension). This may come from anomalies not being far enough in the tail of the normal distribution, *e.g.* forming a cluster near the support of the latter distribution.

Same conclusions and similar accuracies hold for MV-score, which only makes one additional error on the pair (iForest on $pima$, OCSVM on $pima$). Considering all the 36 pairs (9.9), one observes 75% of good comparisons *w.r.t.* ROC-score, and 72% *w.r.t.* PR score. Considering the pairs which are similarly ordered by ROC and PR criteria, this rate increases to $25/33 = 76\%$. The errors are essentially made on *shuttle*, *wild* and *annthyroid* datasets.

Results from the unsupervised framework (training and testing data are polluted by outliers) are similar for both EM and MV criteria. We just observe a slight decrease in accuracy. Considering all the pairs, one observes $26/36 = 72\%$ (resp. $27/36 = 75\%$) of good comparisons *w.r.t.* ROC-score (resp. *w.r.t.* PR score) for EM, and 75% (resp. 78%) of good comparisons *w.r.t.* ROC-score (resp. *w.r.t.* PR score) for MV. Considering the pairs which are similarly ordered by ROC and PR criteria, the rate for EM as for MV increases to $24/31 = 77\%$.

To conclude, 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.

Remark 9.5. (Alternative measurements) There are other ways to measure the accuracy of EM/MV. We can also compute a multi-label classification accuracy, by assigning a label to each algorithm for each experiment (best (B), worst (W), or in-between (I)) according to ROC/PR, and by looking if EM (or MV) is able to recover this label. This methodology is based on the Hamming distance between two rankings, while the previous one is based on the Kendall distance. One drawback of the Hamming distance is that within our setting, the opposite of the ROC score 1 – ROC has a 33% accuracy (the label I is recovered). It has a 0% accuracy with the Kendall distance which counts how many pairs are well-ordered. An other drawback of the Hamming distance to compare rankings is that one single mistake (e.g. an over-ranking of one single algorithm) can shift the labels and yields a 0% accuracy, while the order is globally recovered. (This drawback is limited in our case since we only consider rankings of length three). That said, it is still interesting to have an additional measurement. With this measure, EM has a 87% accuracy in the novelty detection setting, and 59% in the unsupervised one. MV has a 65% accuracy in the novelty detection setting, and 59% in the unsupervised one. According to this way of comparing EM and MV to ROC and PR, EM is preferable to MV in the novelty detection framework. In the unsupervised framework, performance of EM and MV are similar and relatively low.

9.5 Conclusion

One (almost) does not need labels to evaluate anomaly detection algorithms (on continuous data). According to our benchmarks, the excess-mass and mass-volume based numerical criteria introduced in this chapter are (in approximately 80 percent of the cases) able to recover the performance order of algorithms on a fixed dataset (with potentially large dimensionality), without using labels. High-dimensional datasets are dealt with using a method based on feature sub-sampling. This method also brings flexibility to EM and MV criteria, allowing for instance to evaluate the importance of features.

9.6 Further material on the experiments

FIGURE 9.3: ROC and PR curves for Isolation Forest (novelty detection framework)

fig_source/evaluation_fig/bench_iforest_roc_pr_supervised_factorized.png

FIGURE 9.4: ROC and PR curves for Isolation Forest (unsupervised framework)

fig_source/evaluation_fig/bench_iforest_roc_pr_unsupervised_factorized.png

FIGURE 9.5: ROC and PR curves for One Class SVM (novelty detection framework)

fig_source/evaluation_fig/bench_ocsvm_roc_pr_supervised_factorized.png

FIGURE 9.6: ROC and PR curves for One Class SVM (unsupervised framework)

fig_source/evaluation_fig/bench_ocsvm_roc_pr_unsupervised_factorized.png

FIGURE 9.7: ROC and PR curves for Local Outlier Factor (novelty detection framework)

fig_source/evaluation_fig/bench_lof_roc_pr_supervised_factorized.png

FIGURE 9.8: ROC and PR curves for Local Outlier Factor (unsupervised framework)

fig_source/evaluation_fig/bench_lof_roc_pr_unsupervised_factorized.png

FIGURE 9.9: MV and EM curves for http dataset (novelty detection framework)

fig_source/evaluation_fig/mv_em_http_supervised_09_factorized.png

FIGURE 9.10: MV and EM curves for http dataset (unsupervised framework)

fig_source/evaluation_fig/mv_em_http_unsupervised_09_factorized.png

FIGURE 9.11: MV and EM curves for pima dataset (novelty detection framework)

fig_source/evaluation_fig/mv_em_pima_supervised_09_factorized.png

FIGURE 9.12: MV and EM curves for pima dataset (unsupervised framework)

fig_source/evaluation_fig/mv_em_pima_unsupervised_09_factorized.png

FIGURE 9.13: MV and EM curves for smtp dataset (novelty detection framework)

fig_source/evaluation_fig/mv_em_smtp_supervised_09_factorized.png

FIGURE 9.14: MV and EM curves for smtp dataset (unsupervised framework)

fig_source/evaluation_fig/mv_em_smtp_unsupervised_09_factorized.png

FIGURE 9.15: MV and EM curves for wilt dataset (novelty detection framework)

fig_source/evaluation_fig/mv_em_wilt_supervised_09_factorized.png

FIGURE 9.16: MV and EM curves for wilt dataset (unsupervised framework)

fig_source/evaluation_fig/mv_em_wilt_unsupervised_09_factorized.png

FIGURE 9.17: MV and EM curves for adult dataset (novelty detection framework).

fig_source/evaluation_fig/mv_em_adult_supervised_09_factorized.png

FIGURE 9.18: MV and EM curves for adult dataset (unsupervised framework)

fig_source/evaluation_fig/mv_em_adult_unsupervised_09_factorized.png

CHAPTER 10
One Class Splitting Criteria for Random Forests

We recall that this is a contribution of heuristic nature and not yet supported by statistical results. This ongoing work has not been published yet and will certainly be completed in the near future, but we believe that it has its place in our manuscript, given the convincing empirical experiments we carried out and the rationale behind the approach promoted we gave.

Abstract This chapter presents the details relative to the introducing section 1.5.2.
Random Forests (RFs) are strong machine learning tools for classification and regression. However, they remain supervised algorithms, and no extension of RFs to the one-class setting has been proposed, except for techniques based on second-class sampling. This work fills this gap by proposing a natural methodology to extend standard splitting criteria to the one-class setting, structurally generalizing RFs to one-class classification. An extensive benchmark of seven state-of-the-art anomaly detection algorithms is also presented. This empirically demonstrates the relevance of our approach.

Note: The material of this chapter is based on submitted work (Goix et al., 2016a).

10.1 Introduction

Anomaly detection generally aims at finding patterns/observations in data that do not conform to the expected behavior. Anomalies are usually assumed to lie in low probability regions of the data generating process. This assumption drives many statistical anomaly detection methods. Parametric techniques (Barnett & Lewis, 1994; Eskin, 2000) suppose that the normal data are generated by a distribution belonging to some specific parametric model a priori known. Here and hereafter, the term ‘normal data’ does not refer to the Gaussian distributed data, but rather to *not abnormal* ones, *i.e.* data belonging to the above mentioned majority. Classical non-parametric approaches are based on density (level set) estimation (Schölkopf et al., 2001; Scott & Nowak, 2006; Breunig et al., 2000; Quinn & Sugiyama, 2014), on dimensionality reduction (Shyu et al., 2003; Aggarwal & Yu, 2001) or on decision trees (Liu et al., 2008; Shi & Horvath, 2012). Relevant overviews of current research on anomaly detection can be found in Hodge & Austin (2004); Chandola et al. (2009); Patcha & Park (2007); Markou & Singh (2003).

The algorithm proposed in this chapter lies in the *novelty detection* setting, also called *semi-supervised* anomaly detection or *one-class classification*. In this framework, we assume that we only observe examples of one class (referred as the normal class, or inlier class). The second (hidden) class is called the abnormal class, or outlier class. The goal is to identify characteristics of the normal class, such as its support or some density level sets with levels close to zero. This setup is for instance used in some (non-parametric) kernel methods such as One-Class Support Vector Machine algorithm (OCSVM) (Schölkopf et al., 2001), which extends the SVM methodology (Cortes & Vapnik, 1995; Shawe-Taylor & Cristianini, 2004) to handle training using only normal observations (see Section 5.2.1). Recently, Least Squares Anomaly Detection (LSAD) (Quinn & Sugiyama, 2014) similarly extends a multi-class probabilistic classifier (Sugiyama, 2010) to the one-class setting. Both OCSVM and LSAD algorithms extend *structurally* the corresponding classification framework, namely without artificially creating a second class to fall back on a two-class problem. The methodology proposed in this work applies the same structural effort to Random Forests (RFs).

RFs (Breiman, 2001) 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. RFs are strong machine learning tools, comparing well with state-of-the-art methods such as SVM or boosting algorithms (Freund & Schapire, 1996), and used in a wide range of domains (Svetnik et al., 2003; Díaz-Uriarte & De Andres, 2006; Genuer et al., 2010). Practical and theoretical insights on RFs are given in Genuer et al. (2008); Biau et al. (2008); Louppe (2014); Biau & Scornet (2016).

Yet few attempts have been made to transfer the idea of RFs to one-class classification (Désir et al., 2012; Liu et al., 2008; Shi & Horvath, 2012). In Liu et al. (2008), the novel concept of *isolation* is introduced: the Isolation Forest algorithm isolates anomalies, instead of profiling the normal behavior which is the usual approach. It avoids adapting splitting rules to the one-class setting by using extremely randomized trees, also named extra trees (Geurts et al., 2006): isolation trees are built completely randomly, without any splitting rule. Therefore, Isolation Forest is not really based on RFs, the base estimators being extra trees instead of classical decision trees. However, Isolation Forest performs very well in practice with low memory and time complexities. In Désir et al. (2012); Shi & Horvath (2012), outliers are generated to artificially form a second class. In Désir et al. (2012) the authors propose a technique to reduce the number of outliers needed by shrinking the dimension of the input space. The outliers are then generated from the reduced space using a distribution complementary to the 'normal' distribution. Thus their algorithm artificially generates a second class, to use classical RFs. In Shi & Horvath (2012), two different outliers generating processes are compared. In the first one, an artificial second class is added by randomly sampling from the product of empirical marginal distributions. In the second one outliers are uniformly generated from the hyper-rectangle that contains the observed data. The first option is claimed to work best in practice, which can be understood from the curse of dimensionality argument: in large dimension (Tax & Duin, 2002), when the outliers distribution is not tightly defined around the target set, the chance for an outlier to be in the target set becomes very small, so that a huge number of outliers is needed.

Looking beyond the RF literature, Scott & Nowak (2006) propose a methodology to build dyadic decision trees to estimate minimum-volume sets (Polonik, 1997; Einmahl & Mason, 1992). This is done by reformulating their structural risk minimization problem to be able to use Blanchard et al. (2004)'s algorithm. While this methodology can also be used for non-dyadic trees pruning (assuming such a tree has been previously constructed, *e.g.* using some greedy heuristic), it does not allow to effectively grow such trees. A dyadic structure has to be assumed, to be able to 1) link the tree growing process with the general optimization problem and 2) to control the complexity of the resulting partitions, with respect to the regularity of the underlying distribution. In other words, this strong tree structure is needed to derive theoretical guaranties. In the same spirit, Clémençon & Robbiano (2014) proposes to use the two-class splitting criterion defined in Clémençon & Vayatis (2009b). This two-class splitting rule aims at producing oriented decision trees with a 'left-to-right' structure to address the bipartite ranking task. Extension to the one-class setting is done by assuming a uniform distribution for the outlier class. This left-to-right structure is needed to reduce the tree building process to a recursive optimization procedure, thus allowing to derive consistency and rate bounds. Thus, in these two references (Scott & Nowak, 2006; Clémençon & Robbiano, 2014), the priority is given to the theoretical analysis. This imposes constraints on the tree structure which becomes far from the general structure of the base estimators in RF. The price to pay is in the flexibility of the model, and its ability to capture complex broader patterns or structural characteristics from the data.

In this paper, we make the choice to stick to the RF framework. We do not assume any structure for the binary decision trees. The price to pay is the lack of theoretical guaranties, the gain is that we keep the flexibility of RF and are thus able to compete with state-of-the-art anomaly detection algorithms. Besides, we do not assume any (fixed in advance) outlier distribution as in Clémençon & Robbiano (2014), but define it in an adaptive way during the tree building process.

To the best of our knowledge, no algorithm structurally extends (without second class sampling and without alternative base estimators) RFs to one-class classification. Here we precisely introduce 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.

Basic idea. To split a node without second class (outliers) examples, the idea is as follows. Each time we are looking for the best split for a node t, we simply replace (in the two-class *impurity decrease* to be maximized (10.4)) the second class proportion going to the left child node t_L by the proportion expectation $\text{Leb}(t_L)/\text{Leb}(t)$ (idem for the right node), where $\text{Leb}(t)$ denotes the volume of the rectangular cell corresponding to node t. It ensures that one child node tries to capture the maximum number of observations with a minimal volume, while the other child looks for the opposite.

This simple idea corresponds in fact to an adaptive modeling of the outlier distribution. The proportion expectation mentioned above being weighted proportionally to the number of normal instances in node t, the resulting outlier distribution is tightly concentrated around the inliers. Besides, and this attests the consistency of our approach with the two-class framework, it turns out that the one-class model promoted here corresponds to the asymptotic behavior of an adaptive (*w.r.t.* the tree growing process) outliers generating methodology.

This chapter is structured as follows. Section 10.2 provides the reader with necessary background, to address Section 10.3 which proposes a generic adaptation of RFs to the one-class setting and describes a generic one-class random forest algorithm. The latter is compared empirically with state-of-the-art anomaly detection methods in Section 10.4. Finally a theoretical justification of the one-class criterion is given in Section 10.5.

10.2 Background on decision trees

Let us denote by $\mathcal{X} \subset \mathbb{R}^d$ the d-dimensional hyper-rectangle containing all the observations. Consider a binary tree on $\mathcal{X}$ whose node values are subsets of $\mathcal{X}$, iteratively produced by splitting $\mathcal{X}$ into two disjoint subsets. Each internal node t with value $\mathcal{X}_t$ is labeled with a split feature m_t and split value c_t (along that feature), in such a way that it divides $\mathcal{X}_t$ into two disjoint spaces $\mathcal{X}_{t_L} := \{x \in \mathcal{X}_t, x_{m_t} < c_t\}$ and $\mathcal{X}_{t_R} := \{x \in \mathcal{X}_t, x_{m_t} \geq c_t\}$, where t_L (resp. t_R) denotes the left (resp. right) children of node t, and x_j denotes the jth coordinate of vector x. Such a binary tree is grown from a sample $X_1, \dots, X_n$ (for all i, $X_i \in \mathcal{X}$) and its finite depth is determined either by a fixed maximum depth value or by a stopping criterion evaluated on the nodes (*e.g.* based on an impurity measure). The external nodes (or the *leaves*) form a partition of the input space $\mathcal{X}$.

In a supervised classification setting, these binary trees are called *classification trees* and prediction is made by assigning to each sample $x \in \mathcal{X}$ the majority class of the leaves containing x. This is called the *majority vote*. Classification trees are usually built using an impurity measure $i(t)$ whose decrease is maximized at each split of a node t, yielding an optimal split (m_t^*, c_t^*). The decrease of impurity (also called *goodness of split*) $\Delta i(t, t_L, t_R)$ *w.r.t.* the split (m_t, c_t) corresponding to the partition $\mathcal{X}_t = \mathcal{X}_{t_L} \sqcup \mathcal{X}_{t_R}$ of the node t is defined as

$$\Delta i(t, t_L, t_R) = i(t) - p_L i(t_L) - p_R i(t_R), \tag{10.1}$$

where $p_L = p_L(t)$ (resp. $p_R = p_R(t)$) is the proportion of samples from $\mathcal{X}_t$ going to $\mathcal{X}_{t_L}$ (resp. to $\mathcal{X}_{t_R}$). The impurity measure $i(t)$ reflects the goodness of node t: the smaller $i(t)$, the purer the node t and the better the prediction by majority vote on this node. Usual choices for $i(t)$ are the Gini index (Gini, 1912) and the Shannon entropy (Shannon, 2001). To produce a randomized tree, these optimization steps are usually partially randomized (conditionally on the data, splits (m_t^*, c_t^*)'s become random variables), and a classification tree can even be grown totally randomly (Geurts et al., 2006). In a two-class classification setup, the Gini index is

$$i_G(t) = 2\left(\frac{n_t}{n_t + n_t'}\right)\left(\frac{n_t'}{n_t + n_t'}\right) = 1 - \frac{n_t^2 + n_t'^2}{(n_t + n_t')^2}, \tag{10.2}$$

where n_t (resp. n_t') stands for the number of observations with label 0 (resp. 1) in node t. The Gini index is maximal when $n_t/(n_t + n_t') = n_t'/(n_t + n_t') = 0.5$, namely when the conditional probability to have label 0 given that we are in node t is the same as to have label 0 unconditionally: the node t does not discriminate at all between the two classes.

For a node t, maximizing the impurity decrease (10.1) is equivalent to minimizing $p_L i(t_L) + p_R i(t_R)$. As $p_L = (n_{t_L} + n_{t_L}')/(n_t + n_t')$ and $p_R = (n_{t_R} + n_{t_R}')/(n_t + n_t')$, and the quantity $(n_t + n_t')$ being constant in the optimization problem, this is equivalent to minimizing the following proxy of the impurity decrease:

$$I(t_L, t_R) = (n_{t_L} + n_{t_L}')i(t_L) + (n_{t_R} + n_{t_R}')i(t_R). \tag{10.3}$$

Note that if we consider the Gini index as the impurity criteria, the corresponding proxy of the impurity decrease is

$$I_G(t_L, t_R) = \frac{n_{t_L} n_{t_L}'}{n_{t_L} + n_{t_L}'} + \frac{n_{t_R} n_{t_R}'}{n_{t_R} + n_{t_R}'}. \tag{10.4}$$

In the one-class setting, no label is available, hence the impurity measure $i(t)$ does not apply to this setup. The standard splitting criterion which consists in minimizing the latter cannot be used anymore.

10.3 Adaptation to the one-class setting

The two reasons why RFs do not apply to one-class classification are that the standard splitting criterion does not apply to this setup, as well as the majority vote. In this section, we propose a one-class splitting criterion and a natural one-class version of the majority vote.

10.3.1 One-class splitting criterion

As one does not observe the second-class (outliers), n'_t needs to be defined. In the naive approach below, it is defined as $n'_t := n'\mathrm{Leb}(\mathcal{X}_t)/\mathrm{Leb}(\mathcal{X})$, where n' is the supposed total number of (hidden) outliers. In the adaptive approach hereafter, it will be defined as $n'_t := \gamma n_t$, with typically $\gamma = 1$. Thus, the class ratio $\gamma_t := n'_t/n_t$ is defined in both approaches and goes to 0 when $\mathrm{Leb}(\mathcal{X}_t) \to 0$ in the naive approach, while it is maintained constant $\gamma_t \equiv \gamma$ in the adaptive one.

Naive approach. A naive approach to extend the Gini splitting criterion to the one-class setting is to assume a uniform distribution for the second class (outliers), and to replace their number n'_t in node t by the expectation $n'\mathrm{Leb}(\mathcal{X}_t)/\mathrm{Leb}(\mathcal{X})$, where n' denotes the total number of outliers (for instance, it can be chosen as a proportion of the number of inliers). Here and hereafter, Leb denotes the Lebesgue measure on $\mathbb{R}^d$. The problem with this approach appears when the dimension is *not small*. As mentioned in the introduction (curse of dimensionality), when actually generating n' uniform outliers on $\mathcal{X}$, the probability that a node (sufficiently small to yield a good precision) contains at least one of them is very close to zero. That is why data-dependent distributions for the outlier class are often considered (Désir et al., 2012; Shi & Horvath, 2012). Taking the expectation $n'\mathrm{Leb}(\mathcal{X}_t)/\mathrm{Leb}(\mathcal{X})$ instead of the number of points in node t does not solve the curse of dimensionality mentioned in the introduction: the volume proportion $L_t := \mathrm{Leb}(\mathcal{X}_t)/\mathrm{Leb}(\mathcal{X})$ is very close to 0 for nodes t deep in the tree, specially in large dimension. In addition, we typically grow trees on sub-samples of the input data, meaning that even the root node of the trees may be very small compared to the hyper-rectangle containing all the input data. An other problem is that the Gini splitting criterion is skew-sensitive (Flach, 2003), and has here to be apply on nodes t with $0 \simeq n'_t \ll n_t$. When trying empirically this approach, we observe that splitting such nodes produces a child containing (almost) all the data (see Section 10.5).

Remark 10.1. To illustrate the fact that the volume proportion $L_t := \mathrm{Leb}(\mathcal{X}_t)/\mathrm{Leb}(\mathcal{X})$ becomes very close to zero in large dimension for lots of nodes t (in particular the leaves), suppose for the sake of simplicity that the input space is $\mathcal{X} = [0,1]^d$. Suppose that we are looking for a rough precision of $1/2^3 = 0.125$ in each dimension, *i.e.* a unit cube precision of 2^{-3d}. To achieve such a precision, the splitting criterion has to be used on nodes/cells t of volume of order 2^{-3d}, namely with $L_t = 1/2^{3d}$. Note that if we decide to choose n' to be 2^{3d} times larger than the number of inliers in order that $n'L_t$ is not negligible *w.r.t.* the number of inliers, the same (reversed) problem of unbalanced classes appears on nodes with small depth.

Adaptive approach. Our solution is to remove the uniform assumption on the outliers, and to choose their distribution adaptively in such a way it is tightly concentrated around the inlier distribution. Formally, the idea is to maintain constant the class ratio $\gamma_t := n'_t/n_t$ on each node t: before looking for the best split, we update the number of outliers to be equal (up to a scaling constant γ) to the number of inliers, $n'_t = \gamma n_t$, *i.e.* $\gamma_t \equiv \gamma$. These (hidden) outliers are uniformly distributed on node t. The parameter γ is typically set to $\gamma = 1$, see Remark 10.5.

Resulting density. Figure 10.1 shows the corresponding outlier density G. Note that G is a piece-wise approximation of the inlier distribution F. Considering the Neyman-Pearson test $X \sim F$ vs. $X \sim G$ instead of $X \sim F$ vs. $X \sim$ Leb may seem unusual at first sight. However, note that there is $\epsilon > 0$ such that $G > \epsilon$ on the entire input space, since the density

G is constant on each node and equal to the average of F on this node *before splitting it*. If the average of F was estimated to be zero (no inlier in the node), the node would not have been splitted, from where the existence of ϵ. Thus, one might think of G as a piece-wise approximation of $F_\epsilon := (1-\epsilon)F + \epsilon\text{Leb}$. Yet, one can easily show that optimal tests for the Neyman-Pearson problem $H_0 : X \sim F$ vs. $H_1 : X \sim F_\epsilon$ are identical to the optimal tests for $H_0 : X \sim F$ vs. $H_1 : X \sim \text{Leb}$, since the corresponding likelihood ratios are related by a monotone transformation, see Scott & Blanchard (2009) for instance (in fact, this reference shows that these two problems are even equivalent in terms of consistency and rates of convergence of the learning rules).

With this methodology, one cannot derive a one-class version of the Gini index (10.2), but we can define a one-class version of the proxy of the impurity decrease (10.4), by simply replacing n'_{t_L} (resp. n'_{t_R}) by $n'_t\lambda_L$ (resp. $n'_t\lambda_R$), where $\lambda_L := \text{Leb}(\mathcal{X}_{t_L})/\text{Leb}(\mathcal{X}_t)$ and $\lambda_R := \text{Leb}(\mathcal{X}_{t_R})/\text{Leb}(\mathcal{X}_t)$ are the volume proportion of the two child nodes:

$$I_G^{OC-ad}(t_L, t_R) = \frac{n_{t_L}\gamma n_t \lambda_L}{n_{t_L} + \gamma n_t \lambda_L} + \frac{n_{t_R}\gamma n_t \lambda_R}{n_{t_R} + \gamma n_t \lambda_R}. \tag{10.5}$$

Minimization of the one-class Gini improvement proxy (10.5) is illustrated in Figure 10.2. Note that $n'_t\lambda_L$ (resp. $n'_t\lambda_R$) is the expectation of the number of uniform observations (on $\mathcal{X}_t$) among n'_t falling into the left (resp. right) node.

Choosing the split minimizing $I_G^{OC-ad}(t_L, t_R)$ at each step of the tree building process, corresponds to generating $n'_t = \gamma n_t$ outliers each time the best split has to be chosen for node t, and then using the classical two-class Gini proxy (10.4). The only difference is that n'_{t_L} and n'_{t_R} are replaced by their expectations $n'_t\lambda_{t_L}$ and $n'_t\lambda_{t_R}$ in our method.

Remark 10.2. (BY-PRODUCT: EFFICIENTLY GENERATING OUTLIERS) As a by-product, we obtain an efficient method to generate outliers tightly concentrated around the support of the normal distribution: it suffices to generate them as described above, recursively during the tree building process. Sampling n'_t uniform points on $\mathcal{X}_t$, then using the latter to find the best split *w.r.t.* (10.4), and recommence on $\mathcal{X}_{t_L}$ and $\mathcal{X}_{t_R}$.

Remark 10.3. (EXTENSION TO OTHER IMPURITY CRITERIA) Our extension to the one-class setting also applies to other impurity criteria. For instance, in the case of the Shannon entropy defined in the two-class setup by $i_S(t) = \frac{n_t}{n_t+n'_t}\log_2\frac{n_t+n'_t}{n_t} + \frac{n'_t}{n_t+n'_t}\log_2\frac{n_t+n'_t}{n'_t}$, the one-class impurity improvement proxy becomes $I_S^{OC-ad}(t_L, t_R) = n_{t_L}\log_2\frac{n_{t_L}+\gamma n_t\lambda_L}{n_{t_L}} + n_{t_R}\log_2\frac{n_{t_R}+\gamma n_t\lambda_R}{n_{t_R}}$.

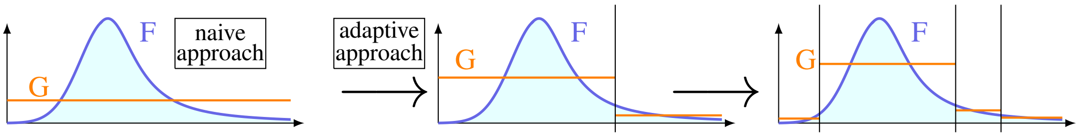


FIGURE 10.1: Outliers distribution G in the naive and adaptive approach. In the naive approach, G does not depends on the tree and is constant on the input space. In the adaptive approach the distribution depends on the inlier distribution F through the tree. The outliers density is constant and equal to the average of F on each node before splitting it.

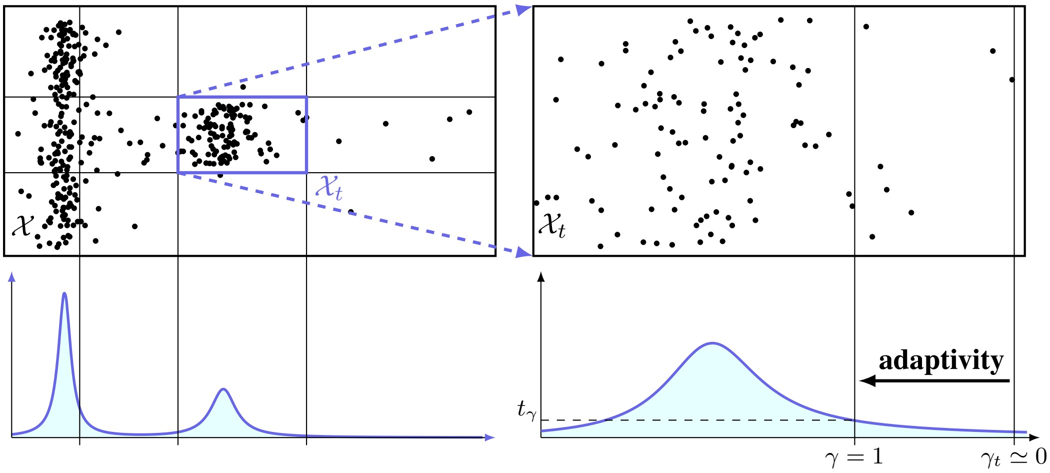


FIGURE 10.2: The left part of this figure represents the dataset under study and the underlying density. After some splits on this initial node $\mathcal{X}$, let us consider the node $\mathcal{X}_t$ illustrated in the right part of this figure: without the proposed adaptive approach, the class ratio γ_t becomes too small and leads to poor splits (all the data are in the 'normal side' of the split, which thus does not discriminate at all). Contrariwise, setting γ to one, *i.e.* using our adaptive approach, is far preferable. Note that a given γ corresponds to a level set t_γ.

10.3.2 Prediction: a majority vote with one single candidate?

Now that RFs can be grown in the one-class setting using our one-class splitting criterion, the forest has to return a prediction adapted to this framework. In other words we also need to extend the concept of majority vote. Most usual one-class (or more generally anomaly detection) algorithms actually provide more than just a level-set estimate or a predicted label for any new observation, abnormal vs. normal. Instead, they return a real valued function, termed *scoring function*, defining a pre-order/ranking on the input space. Such a function $s : \mathbb{R}^d \to \mathbb{R}$ permits to rank any observations according to their supposed 'degree of abnormality'. Thresholding it provides level-set estimates, as well as a decision rule that splits the input space into 'normal' and 'abnormal' regions. The scoring function $s(x)$ we use is the one defined in Liu et al. (2008). It is a decreasing function of the average depth of the leaves containing x in the forest, 'if the trees were fully grown': an average term is added to each node containing more than one sample, say containing N samples. This term $c(N)$ is the average depth of an extremely randomized tree (Geurts et al., 2006) (*i.e.* built without minimizing any criterion, by randomly choosing one feature and one uniform value over this feature to split on) on N samples. Formally,

$$\log_2 s(x) = -\left(\sum_{t \text{ leaves}} \mathbb{1}_{\{x \in t\}} d_t + c(n_t) \right) \ / \ c(n), \tag{10.6}$$

where d_t is the depth of node t, and $c(n) = 2H(n-1) - 2(n-1)/n$, $H(i)$ being the harmonic number.

Remark 10.4 (ALTERNATIVE SCORING FUNCTIONS). Although we use the scoring function defined in (10.6) because of its established high performance (Liu et al., 2008), other scoring functions can be defined. A natural idea to adapt the majority vote to the one-class setting is to change the single vote of a leaf node t into the fraction $\frac{n_t}{\text{Leb}(\mathcal{X}_t)}$, the forest output being the average of the latter quantity over the forest, $s(x) = \sum_{t \text{ leaves}} \mathbb{1}_{\{x \in t\}} \frac{n_t}{\text{Leb}(\mathcal{X}_t)}$. In such a case, each tree of the forest yields a piece-wise density estimate, on its induced partition. The output produced by the forest is then a *step-wise density estimate*. We could also think about the *local density of a typical cell.* For each point x of the input space, it returns the average number of observations in the leaves containing x, divided by the average volume of such leaves. The output of OneClassRF is then the scoring function $s(x) = \big(\sum_{t \text{ leaves}} \mathbb{1}_{\{x \in t\}} n_t\big)\big(\sum_{t \text{ leaves}} \mathbb{1}_{\{x \in t\}} \mathrm{Leb}(\mathcal{X}_t)\big)^{-1}$, where the sums are over each leave of each tree in the forest. This score can be interpreted as the local density of a 'typical' cell (typical among those usually containing x).

10.3.3 OneClassRF: a Generic One-Class Random Forest algorithm

Let us summarize our One Class Random Forest algorithm, based on generic RFs (Breiman, 2001). It has 6 parameters: $max\_samples$, $max\_features\_tree$, $max\_features\_node$, γ, max_depth, n_trees.

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 (10.5) for each split, using parameter γ (recall that $n_t' := \gamma n_t$), until the maximal depth max_depth is achieved. Minimizing (10.5) is done as introduced in Amit & Geman (1997), 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 $s(x)$, the s's being defined in Section 10.3.2.

Figure 10.3 represents the level set of the scoring function produced by OneClassRF, with only one tree (n_trees= 1) of maximal depth max_depth=4, without sub-sampling, and using the Gini-based one-class splitting criterion with $\gamma = 1$.

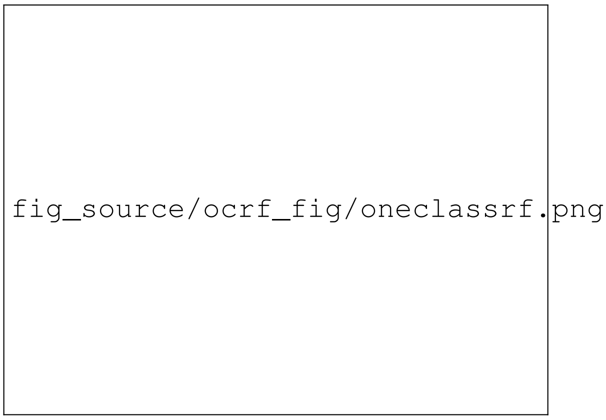


FIGURE 10.3: OneClassRF with one tree: level-sets of the scoring function

Remark 10.5. (INTERPRETATION OF γ) In order for the splitting criterion (10.5) to perform well, n_t' is expected to be of the same order of magnitude as the number of normal observations n_t. If $\gamma = n_t'/n_t \ll 1$, the split puts every normal data on the same side, even the ones which are far in the tail of the distribution, thus widely over-estimating the support of normal data. If $\gamma \gg 1$, the opposite effect happens, yielding an estimate of a t-level set with t not close enough to 0. Figure 10.2 illustrates the splitting criterion when γ varies. It clearly shows that there is a link between parameter γ and the level t_γ of the induced level-set estimate. But from the theory, an explicit relation between γ and t_γ is hard to derive. By default we set γ to 1. One could object that in some situations, it is useful to randomize this parameter. For instance, in the case of a bi-modal distribution for the normal behavior, one split of the tree

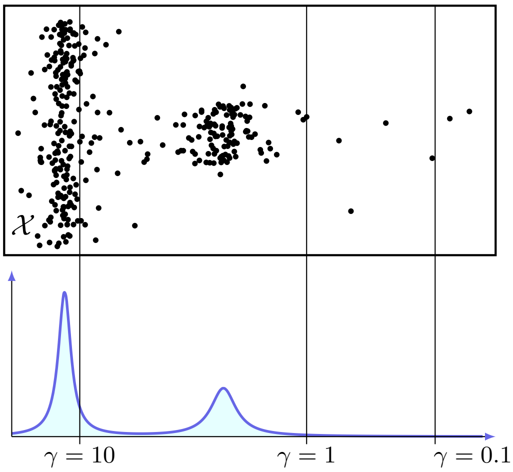


FIGURE 10.4: Illustration of the standard splitting criterion on two modes when the proportion γ varies.

needs to separate two clusters, in order for the level set estimate to distinguish between the two modes. As illustrated in Figure 10.4 , it can only occur if n_t' is large with respect to n_t ($\gamma >> 1$). However, the randomization of γ is somehow included in the randomization of each tree, thanks to the sub-sampling inherent to RFs. Moreover, small clusters tend to vanish when the sub-sample size is sufficiently small: a small sub-sampling size is used in Liu et al. (2008) to isolate outliers even when they form clusters.

Remark 10.6. (ALTERNATIVE STOPPING CRITERIA) Other stopping criteria than a maximal depth may be considered. We could stop splitting a node t when it contains less than n_min observations, or when the quantity $n_t/\text{Leb}(\mathcal{X}_t)$ is large enough (all the points in the cell $\mathcal{X}_t$ are likely to be normal) or close enough to 0 (all the points in the cell $\mathcal{X}_t$ are likely to be abnormal). These options are not discussed in this work.

Remark 10.7. (VARIABLE IMPORTANCE) In the multi-class setting, Breiman (2001) proposed to evaluate the importance of a feature $j \in \{1, \dots d\}$ for prediction by adding up the weighted impurity decreases for all nodes t where X_j is used, averaged over all the trees. The analogue quantity can be computed with respect to the one-class impurity decrease proxy. In our one-class setting, this quantity represents the size of the tail of X_j, and can be interpreted as the capacity of feature j to discriminate between normal/abnormal data.

10.4 Benchmarks

In this section, we compare the OneClassRF algorithm described above to seven state-of-art anomaly detection algorithms: the isolation forest algorithm (Liu et al., 2008) (iForest), a one-class RFs algorithm based on sampling a second-class (Désir et al., 2012) (OCRFsampling), one class SVM (Schölkopf et al., 2001) (OCSVM), local outlier factor (Breunig et al., 2000) (LOF), Orca (Bay & Schwabacher, 2003), Least Squares Anomaly Detection (Quinn & Sugiyama, 2014) (LSAD), Random Forest Clustering (Shi & Horvath, 2012) (RFC).

10.4.1 Default parameters of OneClassRF.

The default parameters taken for our algorithm are the followings. $max\_samples$ is fixed to 20% of the training sample size (with a minimum of 100); $max\_features\_tree$ is fixed to

TABLE 10.1: Original Datasets characteristics

Datasets	nb of samples	nb of features	anomaly class	
adult	48842	6	class '$> 50K$'	(23.9%)
annthyroid	7200	6	classes $\neq 3$	(7.42%)
arrhythmia	452	164	classes $\neq 1$ (features 10-14 removed)	(45.8%)
forestcover	286048	10	class 4 (vs. class 2)	(0.96%)
http	567498	3	attack	(0.39%)
ionosphere	351	32	bad	(35.9%)
pendigits	10992	16	class 4	(10.4%)
pima	768	8	pos (class 1)	(34.9%)
shuttle	85849	9	classes $\neq 1$ (class 4 removed)	(7.17%)
smtp	95156	3	attack	(0.03%)
spambase	4601	57	spam	(39.4%)
wilt	4839	5	class 'w' (diseased trees)	(5.39%)

TABLE 10.2: Results for the novelty detection setting (semi-supervised framework). The table reports AUC ROC and AUC PR scores (higher is better) for each algorithms. The training time of each algorithm has been limited (for each experiment among the 10 performed for each dataset) to 30 minutes, where 'NA' indicates that the algorithm could not finish training within the allowed time limit. In average on all the datasets, our proposed algorithm 'OneClassRF' achieves both best AUC ROC and AUC PR scores (with LSAD for AUC ROC). It also achieves the lowest cumulative training time.

Dataset	OneClassRF		iForest		OCRFsampl.		OCSVM		LOF		Orca		LSAD		RFC	
	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR
adult	**0.665**	**0.278**	0.661	0.227	NA	NA	0.638	0.201	0.615	0.188	0.606	0.218	0.647	0.258	NA	NA
annthyroid	**0.936**	0.468	0.913	0.456	0.918	**0.532**	0.706	0.242	0.832	0.446	0.587	0.181	0.810	0.327	NA	NA
arrhythmia	0.684	0.510	0.763	0.492	0.639	0.249	**0.922**	**0.639**	0.761	0.473	0.720	0.466	0.778	0.514	0.716	0.299
forestcover	0.968	0.457	0.863	0.046	NA	NA	NA	NA	**0.990**	**0.795**	0.946	0.558	0.952	0.166	NA	NA
http	**0.999**	**0.838**	0.994	0.197	NA	NA	NA	NA	NA	NA	**0.999**	0.812	0.981	0.537	NA	NA
ionosphere	0.909	0.643	0.902	0.535	0.859	0.609	0.973	0.849	0.959	0.807	0.928	**0.910**	**0.978**	0.893	0.950	0.754
pendigits	0.960	0.559	0.810	0.197	0.968	0.694	0.603	0.110	0.983	0.827	**0.993**	**0.925**	0.983	0.752	NA	NA
pima	0.719	0.247	0.726	0.183	**0.759**	**0.266**	0.716	0.237	0.700	0.152	0.588	0.175	0.713	0.216	0.506	0.090
shuttle	**0.999**	**0.998**	0.996	0.973	NA	NA	0.992	0.924	**0.999**	0.995	0.890	0.782	0.996	0.956	NA	NA
smtp	0.922	0.499	0.907	0.005	NA	NA	0.881	**0.656**	**0.924**	0.149	0.782	0.142	0.877	0.381	NA	NA
spambase	**0.850**	0.373	0.824	0.372	0.797	**0.485**	0.737	0.208	0.746	0.160	0.631	0.252	0.806	0.330	0.723	0.151
wilt	0.593	0.070	0.491	0.045	0.442	0.038	0.323	0.036	0.697	0.092	0.441	0.030	0.677	0.074	**0.896**	**0.631**
average:	**0.850**	**0.495**	0.821	0.311	0.769	0.410	0.749	0.410	0.837	0.462	0.759	0.454	**0.850**	0.450	0.758	0.385
cum. train time:	**61s**		68s		NA		NA		NA		2232s		73s		NA	

50% of the total number of features with a minimum of 5 (*i.e.* each tree is built on 50% of the total number of features); $max\_features\_node$ is fixed to 5; γ is fixed to 1 (see Remark 10.5); max_depth is fixed to $\log_2$ (logarithm in base 2) of the training sample size as in Liu et al. (2008); n_trees is fixed to 100 as in the previous reference; and parameter s_i is set to s_3 as defined in (10.6).

10.4.2 Hyper-Parameters of tested algorithms

Overall we chose to train the different algorithms with their (default) hyper-parameters as seen in their respective paper or author's implementation.

The *OCSVM* algorithm uses default parameters: `kernel='rbf'`, `tol=1e-3`, `nu=0.5`, `shrinking=True`, `gamma=1/n_features`, where tol is the tolerance for stopping criterion.

The *LOF* algorithm uses default parameters: `n_neighbors=5`, `leaf_size=30`, `metric='minkowski'`, `contamination=0.1`, `algorithm='auto'`, where the algorithm parameters stipulates how to compute the nearest neighbors (either ball-tree, kd-tree or brute-force).

The *iForest* algorithm uses default parameters: `n_estimators=100`, `max_samples=min(256, n_samples)`, `max_features=1`, `bootstrap=false`, where bootstrap states whether samples are drawn with replacement.

The *OCRFsampling* algorithm uses default parameters: the number of dimensions for the Random Subspace Method `krsm=-1`, the number of features randomly selected at each node during the induction of the tree `krfs=-1`, `n_tree=100`, the factor controlling the extension of the outlier domain used for outlier generation according to the volume of the hyper-box surrounding the target data `alpha=1.2`, the factor controlling the number of outlier data generated according to the number of target data `beta=10`, whether outliers are generated from uniform distribution `optimize=0`, whether data outside target bounds are considered as outlier data `rejectOutOfBounds=0`.

The *Orca* algorithm uses default parameter `k=5` (number of nearest neighbors) as well as `N=n/8` (how many anomalies are to be reported). The last setting, set up in the empirical evaluation of iForest in Liu et al. (2012), allows a better computation time without impacting Orca's performance.

The *RFC* algorithm uses default parameters: `no.forests=25`, `no.trees=3000`, the Addcl1 Random Forest dissimilarity `addcl1=T, addcl2=F`, use the importance measure `imp=T`, the data generating process `oob.prox1=T`, the number of features sampled at each split `mtry1=3`.

The *LSAD* algorithm uses default parameters: the maximum number of samples per kernel `n_kernels_max=500`, the center of each kernel (the center of the random sample subset by default) `kernel_pos='None'`, the kernel scale parameter (using the pairwise median trick by default)`gamma='None'`, the regularization parameter `rho=0.1`.

10.4.3 Description of the datasets

The characteristics of the twelve reference datasets considered here are summarized in Table 10.1. They are all available on the UCI repository (Lichman, 2013) and the preprocessing is done in a classical way. We removed all non-continuous attributes as well as attributes taking less than 10 different values. The *http* and *smtp* datasets belong to the KDD Cup '99 dataset (KDDCup, 1999; Tavallaee et al., 2009), which consist of a wide variety of hand-injected attacks (anomalies) in a closed network (normal background). They are classically obtained as described in Yamanishi et al. (2000). This two datasets are available on the *scikit-learn* library (Pedregosa et al., 2011). The *shuttle* dataset is the fusion of the training and testing datasets available in the UCI repository. As in Liu et al. (2008), we use instances from all different classes but class 4. In the *forestcover* data, the normal data are the instances from class 2 while instances from class 4 are anomalies (as in Liu et al. (2008)). The *ionosphere* dataset differentiates 'good' from 'bad' radars, considered here as abnormal. A 'good' radar shows evidence of some type of structure in the ionosphere. A 'bad' radar does not, its signal passing through the ionosphere. The *spambase* dataset consists of spam or non-spam emails. The former constitute our anomaly class. The *annthyroid* medical dataset on hypothyroidism contains one normal class and two abnormal ones, which form our outliers. The *arrhythmia* dataset reflects the presence and absence (class 1) of cardiac arrhythmia. The number of attributes being large considering the sample size, we removed attributes containing missing data. The *pendigits* dataset contains 10 classes corresponding to the digits from 0 to 9, examples being handwriting samples. As in Schubert et al. (2012), the abnormal data are chosen to be those from class 4. The *pima* dataset consists of medical data on diabetes. Patients suffering from diabetes (normal class) were considered outliers. The *wild* dataset involves detecting diseased trees in Quickbird imagery. Diseased trees (class 'w') is our abnormal class. In the *adult* dataset, the goal is to predict whether income exceeds $ 50K/year based on census data. We only keep the 6 continuous attributes.

10.4.4 Results

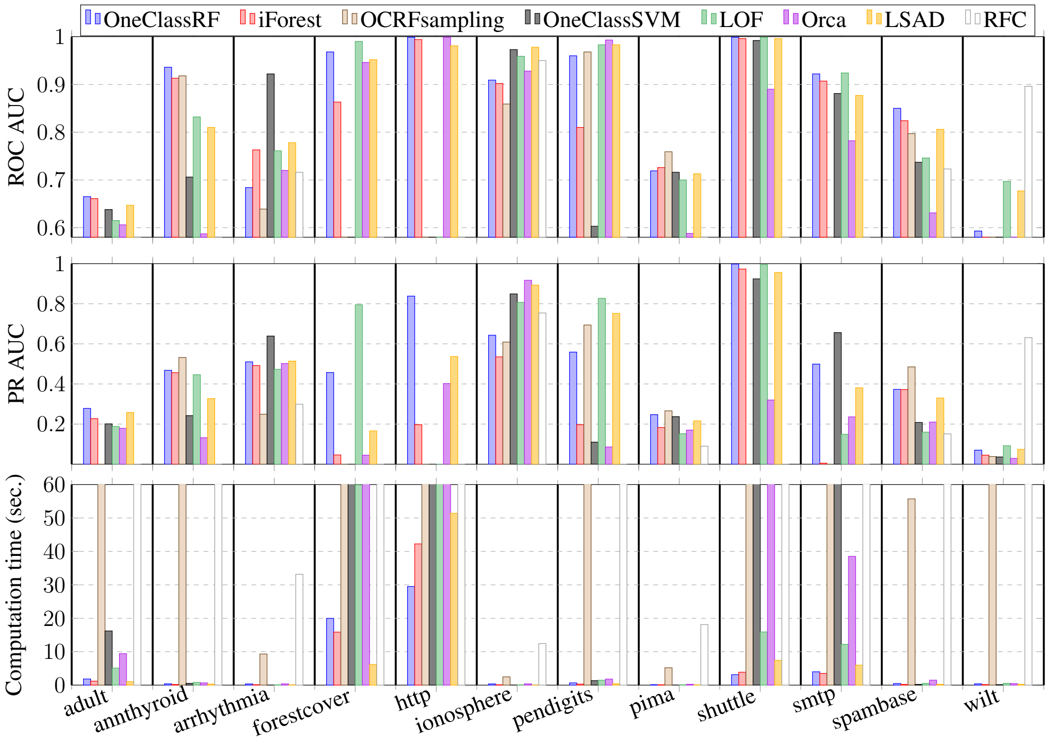


FIGURE 10.5: Performances of the algorithms on each dataset in the novelty detection framework: ROC AUCs are displayed on the top, Precision-Recall AUCs in the middle and training times[1]on the bottom, for each dataset and algorithm. The x-axis represents the datasets.

The experiments are performed in the novelty detection framework, also named semi-supervised anomaly detection, where the training set consists of normal data only. We simply removed anomalies from the training data. For each algorithm, 10 experiments on random training and testing datasets are performed, yielding averaged ROC and Precision-Recall curves whose AUC are summarized in Table 10.2 (see the last section of this chapter to further insights on the benchmarks). It appears that OneClassRF has the best performance on five datasets in terms of ROC AUCs, and is also the best in average. Computation times (training plus testing) of OneClassRF are also very competitive. Figure 10.5 shows that the amount of time to train and test any dataset takes less than one minute with OneClassRF, whereas some algorithms have far higher computation times (OCRFsampling, OneClassSVM, LOF and Orca have computation times higher than 30 minutes in some datasets). Our approach leads to results similar to quite new algorithms such as iForest and LSDA.

[1] For OCRF, Orca and RFC, testing and training time cannot be isolated because of algorithms implementation: for these algorithms, the sum of the training and testing times are displayed in Figure 10.5 and 10.6.

Experiments in an unsupervised framework (the training set is polluted by abnormal data) have also been made. The anomaly rate is arbitrarily bounded to 10% max (before splitting data into training and testing sets).

10.5 Theoretical justification for the one-class splitting criterion

10.5.1 Underlying model

In order to generalize the two-class framework to the one-class one, we need to consider the population versions associated to empirical quantities (10.1), (10.2) and (10.3), as well as the underlying model assumption. The latter can be described as follows.

Existing Two-Class Model (n, $\boldsymbol{\alpha}$). We consider a random variable (*r.v.*) $X : \Omega \to \mathbb{R}^d$ *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. To summarize:

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

Then, considering $p(t_L|t) = \mathbb{P}(X \in \mathcal{X}_{t_L} | X \in \mathcal{X}_t)$, $p(t_R|t) = \mathbb{P}(X \in \mathcal{X}_{t_R} | X \in \mathcal{X}_t)$, the probabilistic version of (10.1) is

$$\Delta i^{theo}(t, t_L, t_R) \;\; = \;\; i^{theo}(t) \; - \; p(t_L|t)\, i^{theo}(t_L) \; - \; p(t_R|t)\, i^{theo}(t_R), \tag{10.7}$$

for instance using the Gini index $i^{theo} = i_G^{theo}$,

$$i_G^{theo}(t) \;\; = \;\; 2\mathbb{P}(y=0|X \in \mathcal{X}_t) \cdot \mathbb{P}(y = 1 | X \in \mathcal{X}_t) \;\; = \;\; \frac{\mathbb{P}(X \in \mathcal{X}_t,\; y = 0) \cdot \mathbb{P}(X \in \mathcal{X}_t,\; y = 1)}{\mathbb{P}(X \in \mathcal{X}_t)^2} \tag{10.8}$$

which is the population version of (10.2). Indeed, when observing n $i.i.d.$ realizations $(X_1, y_1), \dots, (X_n, y_n)$ of (X, y), replacing probabilities by their empirical version amounts to replacing $\mathbb{P}(X \in \mathcal{X}_t,\; y = 0)$ by n_t/n, $\mathbb{P}(X \in \mathcal{X}_t,\; y = 1)$ by n_t'/n and $\mathbb{P}(X \in \mathcal{X}_t)$ by $(n_t + n_t')/n$ with $n_t = \text{card}\{i,\; X_i \in \mathcal{X}_t, y_i = 0\}$ and $n_t' = \text{card}\{i,\; X_i \in \mathcal{X}_t, y_i = 1\}$, thus recovering (10.2).

One-Class-Model ($\boldsymbol{n}$, $\boldsymbol{\alpha}$). 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 $\mathbf{N} \sim \text{Bin}\big(n, (1-\alpha)\big)$. Here and hereafter, $\text{Bin}(n,p)$ denotes the binomial distribution with parameters (n,p). As outliers are not observed, it is natural to 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 $\mathcal{X}$. Note that this assumption *will be removed* in the adaptive approach described below (which aims at maintaining a non-negligible proportion of (hidden) outliers in every nodes).

Let us define $L_t = \text{Leb}(\mathcal{X}_t)/\text{Leb}(\mathcal{X})$. Then, $\mathbb{P}(X \in \mathcal{X}_t,\; y = 1) = \mathbb{P}(y = 1)\mathbb{P}(X \in \mathcal{X}_t |\, y = 1) = \alpha L_t$. Replacing in (10.8) the probability $\mathbb{P}(X \in \mathcal{X}_t, y = 0)$ by its empirical version

n_t/n, we then obtain the one-class empirical Gini index

$$i_G^{OC}(t) \;=\; \frac{n_t \alpha n L_t}{(n_t+\alpha n L_t)^2}. \tag{10.9}$$

In the following, we say that this one-class index is a *semi-empirical* version of (10.8), in the sense that it is obtained by considering empirical quantities for the (observed) normal behavior and population quantities for the (non-observed) abnormal behavior. Now, maximizing the population version of the impurity decrease $\Delta i_G^{theo}(t,t_L,t_R)$ as defined in (10.7) is equivalent to minimizing

$$p(t_L|t)\; i_G^{theo}(t_L) \;+\; p(t_R|t)\; i_G^{theo}(t_R). \tag{10.10}$$

Considering semi-empirical versions of $p(t_L|t)$ and $p(t_R|t)$, as for (10.9), gives $p_n(t_L|t) = (n_{t_L}+\alpha n L_{t_L})/(n_t+\alpha n L_t)$ and $p_n(t_R|t) = (n_{t_R}+\alpha n L_{t_R})/(n_t+\alpha n L_t)$. Then, the semi-empirical version of (10.10) is

$$p_n(t_L|t)\; i_G^{OC}(t_L) \;+\; p_n(t_R|t)\; i_G^{OC}(t_R) = \frac{1}{(n_t+\alpha n L_t)}\left(\frac{n_{t_L}\alpha n L_{t_L}}{n_{t_L}+\alpha n L_{t_L}} + \frac{n_{t_R}\alpha n L_{t_R}}{n_{t_R}+\alpha n L_{t_R}}\right) \tag{10.11}$$

where $1/(n_t+\alpha n L_t)$ is constant when the split varies. This means that finding the split minimizing (10.11) is equivalent to finding the split minimizing

$$I_G^{OC}(t_L,t_R) = \frac{n_{t_L}\alpha n L_{t_L}}{n_{t_L}+\alpha n L_{t_L}} + \frac{n_{t_R}\alpha n L_{t_R}}{n_{t_R}+\alpha n L_{t_R}}. \tag{10.12}$$

Remark 10.8. (DIRECT LINK WITH THE TWO-CLASS FRAMEWORK) Note that the two-class proxy of the Gini impurity decrease (10.4) is easily recovered by replacing $\alpha n L_{t_L}$ (resp. $\alpha n L_{t_R}$) by n'_{t_L} (resp. n'_{t_R}), the number of second class instances in t_L (resp. in t_R). When generating αn of them uniformly on $\mathcal{X}$, $\alpha n L_t$ is the expectation of n'_t.

As detailed Section 10.3.1, this approach suffers from the curse of dimensionality. We can summarize the problem as follows. Note that γ_t, the ratio between the expected number of (hidden) outliers and the number of normal observations in node t, is here equal to

$$\gamma_t = \frac{\alpha n L_t}{n_t}. \tag{10.13}$$

This class ratio is close to 0 for lots of nodes t, which makes unable the Gini criterion to discriminate accurately between the (hidden) outliers and the inliers. Minimizing this criterion produces splits corresponding to $\gamma_t \simeq 0$ in Figure 10.2: one of the two child nodes, say t_L contains almost all the data.

10.5.2 Adaptive approach

The solution presented Section 10.3 is to remove the uniform assumption for the abnormal class. From the theoretical point of view, the idea is to choose in an adaptive way (*w.r.t.* the volume of $\mathcal{X}_t$) the number αn, which can be interpreted as the number of (hidden) outliers. Recall that neither n nor α is observed in the One-Class-Model(n, α). Doing so, we aim at avoiding $\alpha n L_t \ll n_t$ when L_t is too small. Namely, with γ_t defined in (10.13), we aim at avoiding $\gamma_t \simeq 0$ when $L_t \simeq 0$. The idea is to consider $\alpha(L_t)$ and $n(L_t)$ such that $\alpha(L_t) \to 1$, $n(L_t) \to \infty$ when $L_t \to 0$. We then define the one-class adaptive proxy of the impurity decrease by

$$I_G^{OC-ad}(t_L, t_R) = \frac{n_{t_L}\alpha(L_t)\cdot n(L_t)\cdot L_{t_L}}{n_{t_L}+\alpha(L_t)\cdot n(L_t)\cdot L_{t_L}} + \frac{n_{t_R}\alpha(L_t)\cdot n(L_t)\cdot L_{t_R}}{n_{t_R}+\alpha(L_t)\cdot n(L_t)\cdot L_{t_R}}. \tag{10.14}$$

In other words, instead of considering one general model One-Class-Model(n, α) defined in Section 10.5.1, we adapt it to each node t, considering One-Class-Model($n(L_t)$, $\alpha(L_t)$) *before searching the best split.* We still consider the N normal observations as a realization of this model. When growing the tree, using One-Class-Model($n(L_t)$, $\alpha(L_t)$) as L_t becomes close to zero allows to maintain a high expected proportion of outliers in the node to be split minimizing (10.14). Of course, constraints have to be imposed to ensure consistency between these models. 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 $\big(n(L_t), \alpha(L_t)\big)$ is

$$(1-\alpha)n = \big(1-\alpha(L_t)\big)\cdot n(L_t) \quad \text{for all } \ t, \tag{10.15}$$

so that the expectation of $\mathbf{N}$ remains unchanged.

Remark 10.9. In our adaptive model One-Class-Model($n(L_t)$, $\alpha(L_t)$) which varies when we grow the tree, let us denote by $\mathbf{N}(L_t) \sim \text{Bin}\big(n(L_t), 1-\alpha(L_t)\big)$ the *r.v.* ruling the number of normal data. The number of normal observations N is still viewed as a realization of it. Note that the distribution of $\mathbf{N}(L_t)$ converges in distribution to $\mathcal{P}\big((1-\alpha)n\big)$ a Poisson distribution with parameter $(1-\alpha)n$ when $L_t \to 0$, while the distribution $\text{Bin}\big(n(L_t), \alpha(L_t)\big)$ of the *r.v.* $n(L_t) - \mathbf{N}(L_t)$ ruling the number of (hidden) outliers goes to infinity almost surely. In other words, the asymptotic model (when $L_t \to 0$) consists in assuming that the number of normal data N we observed is a realization of $\mathbf{N}_\infty \sim \mathcal{P}\big((1-\alpha)n\big)$, and that an infinite number of outliers have been hidden.

A second natural constraint on $\big(\alpha(L_t), n(L_t)\big)$ concerns on γ_t defined in (10.13), the ratio between the expected number of (hidden) outliers in node t and the number of normal observations. As explained in Section 10.3.1, we do not want γ_t to go to zero when L_t does. Let us say we want γ_t to be constant for all node t, equal to $\gamma > 0$. Typically, $\gamma = 1$ so that there is as much expected uniform (hidden) outliers than normal data at each time we want to find the best split minimizing (10.14). Then the second constraint is

$$\alpha(L_t)\cdot n(L_t)\cdot L_t = \gamma_t n_t = \gamma n_t := n_t'. \tag{10.16}$$

The quantity n_t' can be interpreted as the expected number of (hidden) outliers in node t. The constant γ is a parameter ruling the expected proportion of outliers in each node. Equations (10.15) and (10.16) allow to explicitly determine $\alpha(L_t)$ and $n(L_t)$: $\alpha(L_t) = n_t'/\big((1-\alpha)nL_t + n_t'\big)$ and $n(L_t) = \big((1-\alpha)nL_t + n_t'\big)/L_t$. Regarding (10.14), $\alpha(L_t)\cdot n(L_t)\cdot L_{t_L} = \frac{n_t'}{L_t}L_{t_L} = n_t'\text{Leb}(\mathcal{X}_{t_L})/\text{Leb}(\mathcal{X}_t)$ by (10.16) and $\alpha(L_t)\cdot n(L_t)\cdot L_{t_R} = n_t'\text{Leb}(\mathcal{X}_{t_R})/\text{Leb}(\mathcal{X}_t)$, so that we recover equation (10.5).

10.6 Conclusion

Through a natural adaptation of both (two-class) splitting criteria and majority vote, this chapter introduces a methodology to structurally extend RFs to the one-class setting. Our one-class splitting criteria correspond to the asymptotic behavior of an adaptive outliers generating methodology, so that consistency with two-class RFs seems respected. Strong empirical performance attests the relevance of this methodology.

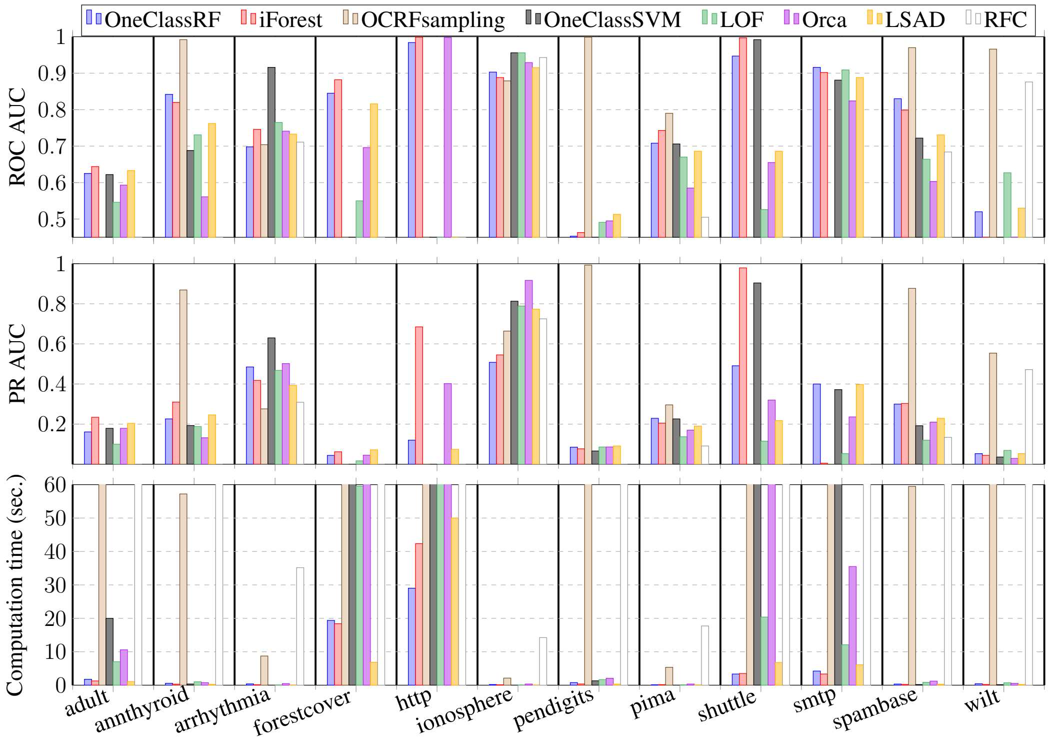


FIGURE 10.6: Performances of the algorithms on each dataset in the unsupervised framework: ROC AUCs are on the top, Precision-Recall AUCs in the middle and processing times are displayed below (for each dataset and algorithm). The x-axis represents the datasets.

10.7 Further details on benchmarks and unsupervised results

Recall that for each algorithm, 10 experiments on random training and testing datasets are performed. Averaged ROC and Precision-Recall curves AUC are summarized in Table 10.2. For the experiments made in an unsupervised framework (meaning that the training set is polluted by abnormal data), the anomaly rate is arbitrarily bounded to 10% max (before splitting data into training and testing sets).

TABLE 10.3: Results for the unsupervised setting

Dataset	OneClassRF		iForest		OCRFsampling		OCSVM		LOF		Orca		LSDA		RFC	
	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR
adult	0.625	0.161	**0.644**	0.234	NA	NA	0.622	0.179	0.546	0.100	0.593	0.179	0.633	0.204	NA	NA
annthyroid	0.842	0.226	0.820	0.310	**0.992**	0.869	0.688	0.193	0.731	0.188	0.561	0.132	0.762	0.246	NA	NA
arrhythmia	0.698	0.485	0.746	0.418	0.704	0.276	**0.916**	0.630	0.765	0.468	0.741	0.502	0.733	0.393	0.711	0.309
forestcover	0.845	0.044	**0.882**	0.062	NA	NA	NA	NA	0.550	0.017	0.696	0.045	0.816	0.072	NA	NA
http	0.984	0.120	**0.999**	0.685	NA	NA	NA	NA	NA	NA	0.998	0.402	0.277	0.074	NA	NA
ionosphere	0.903	0.508	0.888	0.545	0.879	0.664	**0.956**	0.813	**0.956**	0.789	0.929	0.917	0.915	0.773	0.943	0.725
pendigits	0.453	0.085	0.463	0.077	**0.999**	0.993	0.366	0.066	0.491	0.086	0.495	0.086	0.513	0.091	NA	NA
pima	0.708	0.229	0.743	0.205	**0.790**	0.296	0.706	0.226	0.670	0.137	0.585	0.170	0.686	0.190	0.505	0.091
shuttle	0.947	0.491	**0.997**	0.979	NA	NA	0.992	0.904	0.526	0.115	0.655	0.320	0.686	0.218	NA	NA
smtp	**0.916**	0.400	0.902	0.005	NA	NA	0.881	0.372	0.909	0.053	0.824	0.236	0.888	0.398	NA	NA
spambase	0.830	0.300	0.799	0.303	**0.970**	0.877	0.722	0.192	0.664	0.120	0.603	0.210	0.731	0.229	0.684	0.134
wilt	0.520	0.053	0.443	0.044	**0.966**	0.554	0.316	0.036	0.627	0.069	0.441	0.029	0.530	0.053	0.876	0.472
average:	0.773	0.259	0.777	0.322	**0.900**	0.647	0.717	0.361	0.676	0.195	0.677	0.269	0.681	0.245	0.744	0.346
cum. train time:	**61s**		70s		NA		NA		NA		2432s		72s		NA	

ROC and PR curves:

FIGURE 10.7: ROC and PR curves for OneClassRF (novelty detection framework)

fig_source/ocrf_fig/bench_oneclassrf_roc_pr_supervised_factorized.png

FIGURE 10.8: ROC and PR curves for OneClassRF (unsupervised framework)

fig_source/ocrf_fig/bench_oneclassrf_roc_pr_unsupervised_factorized.png

FIGURE 10.9: ROC and PR curves for IsolationForest (novelty detection framework)

fig_source/ocrf_fig/bench_iforest_roc_pr_supervised_factorized.png

FIGURE 10.10: ROC and PR curves for IsolationForest (unsupervised framework)

fig_source/ocrf_fig/bench_iforest_roc_pr_unsupervised_factorized.png

FIGURE 10.11: ROC and PR curves for OCRFsampling (novelty detection framework)

fig_source/ocrf_fig/bench_ocrf_roc_pr_supervised_factorized.png

FIGURE 10.12: ROC and PR curves for OCRFsampling (unsupervised framework)

fig_source/ocrf_fig/bench_ocrf_roc_pr_unsupervised_factorized.png

FIGURE 10.13: ROC and PR curves for OCSVM (novelty detection framework)

fig_source/ocrf_fig/bench_ocsvm_roc_pr_supervised_factorized.png

FIGURE 10.14: ROC and PR curves for OCSVM (unsupervised framework)

fig_source/ocrf_fig/bench_ocsvm_roc_pr_unsupervised_factorized.png

FIGURE 10.15: ROC and PR curves for LOF (novelty detection framework)

fig_source/ocrf_fig/bench_lof_roc_pr_supervised_factorized.png

FIGURE 10.16: ROC and PR curves for LOF (unsupervised framework)

fig_source/ocrf_fig/bench_lof_roc_pr_unsupervised_factorized.png

FIGURE 10.17: ROC and PR curves for Orca (novelty detection framework)

fig_source/ocrf_fig/bench_orca_roc_pr_supervised_factorized.png

FIGURE 10.18: ROC and PR curves for Orca (unsupervised framework)

fig_source/ocrf_fig/bench_orca_roc_pr_unsupervised_factorized.png

FIGURE 10.19: ROC and PR curves for LSAD (novelty detection framework)

fig_source/ocrf_fig/bench_LSAnomaly_roc_pr_supervised_factorized.png

FIGURE 10.20: ROC and PR curves for LSAD (unsupervised framework)

fig_source/ocrf_fig/bench_LSAnomaly_roc_pr_unsupervised_factorized.png

FIGURE 10.21: ROC and PR curves for RFC (novelty detection framework)

fig_source/ocrf_fig/bench_rf_roc_pr_supervised_factorized.png

FIGURE 10.22: ROC and PR curves for RFC (unsupervised framework)

fig_source/ocrf_fig/bench_rf_roc_pr_unsupervised_factorized.png

CHAPTER 11
Conclusion, limitations & perspectives

In this thesis, three different problems have been addressed. Building scoring functions, gaining in accuracy on low probability regions, and evaluating algorithms in the case of unlabeled data.

For the first problem, two solutions have been proposed. The excess-mass based performance criterion has been defined and used in a empirical risk minimization framework. While theoretical guaranties have been derived (resp. exists) for scoring functions produced by optimizing the excess-mass (resp. mass-volume) curves, more work is needed for them to achieve efficiency in practice. In particular, the choice of the functional class on which the criterion is optimized (which is typically the class of stepwise functions on very simple sets) is challenging. This class has to be rich enough to provide a good approximation while simple enough to control the convergence rate and the algorithmic complexity, which is a hard trade-off to achieve when no information on the underlying distribution is used. The second solution is based on random forests. It consists in extending naturally standard splitting criteria to the one-class setting. This structural generalization of random forests to one-class classification produces competitive scoring functions with respect to many state-of-the-art anomaly detection algorithms commonly used in industrial setups. Its principal limitations, which also are perspectives, lie in the fact that this is essentially a heuristic method, and it lacks of theoretical guaranties.

For the accuracy gain on low probability regions, tools borrowed from multivariate extreme value theory have been used to defined a possibly sparse representation of the dependence structure of extremes, and the associated scoring functions. Besides, non-asymptotic bounds have been derived on the estimation procedure. An intermediate step was to study the non-asymptotic behavior of the stable tail dependence function, a functional characterizing the extreme dependence structure. Novel bounds have been derived to control the error of its natural empirical version, as well as a methodology for deriving VC-type bounds on low-probability regions. Moreover, the sparsity pattern in multivariate extremes we exhibit 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. Because no assumption is made on the underlying distribution other than the existence of the STDF, the non-asymptotic bounds on the estimation procedure contain separated bias terms corresponding to the (distribution-dependent) convergence speed to the asymptotic behavior, which are not controlled explicitly. This prevents us to choose in an adaptive way the parameters of our representation of extreme dependence. Since these parameters cannot be chosen by cross-validation as no labeled data is available in our setting, default parameters have to be chosen. While results seem accurate in our benchmarks, this is a limitation of this approach. 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.

For the evaluation of anomaly detection algorithms in the case of unlabeled data, the theory is built on the excess-mass and mass-volume curves. As these criterion have originally been introduced to build scoring functions using empirical risk minimization, no empirical study has been made in the literature on their ability to discriminate between existing scoring functions. To this end, we present a benchmark showing that EM and MV criteria are able to recover the ranking induced by ROC/PR over scoring functions, when labels are hidden. Besides, as these curves can only be estimated in small dimensions (they involve volume computations), a methodology based on feature sub-sampling and aggregating is described and tested. It allows an extension of the use of these criteria to high-dimensional datasets. It also solves major drawbacks inherent to standard EM and MV curves, allowing *e.g.* features selection in the unsupervised setting. This heuristic (random projections and aggregating) works well in practice, but suffers from a lack of theoretical guaranties. Note that it does not evaluate a fixed scoring function, but multiple scoring functions (output by the algorithm to be evaluated) defined on different sub-spaces with the same dimension. Thus, it could be of great interest to study if the quantity estimated here really converges when the number of random projections go to infinity, and to find out if the limit somehow corresponds to a scoring function defined on the entire input space. In particular, this may allow to choose in an optimal way the sub-sampling parameters.

Bibliography

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

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Comput.*, 9:1545–1588, 1997.

M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Appl Math*, 47:207 – 217, 1993.

A. Baillo. Total error in a plug-in estimator of level sets. *Statistics & probability letters*, 65:411–417, 2003.

A. Baillo, J. A Cuesta-Albertos, and A. Cuevas. Convergence rates in nonparametric estimation of level sets. *Statistics & probability letters*, 53:27–35, 2001.

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

E. N Barron, P. Cardaliaguet, and R. Jensen. Conditional essential suprema with applications. *Applied Mathematics and Optimization*, 48:229–253, 2003.

S. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proc. of KDD*, pages 29–38, 2003.

J. Beirlant, M. Escobar-Bach, Y. Goegebeur, and A. Guillou. Bias-corrected estimation of stable tail dependence function. *JMVA*, 143:453–466, 2016.

J. Beirlant and Y. Goegebeur. Local polynomial maximum likelihood estimation for pareto-type distributions. *JMVA*, 89:97–118, 2004.

J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.

J. Beirlant, P. Vynckier, and J. L. Teugels. Tail index estimation, pareto quantile plots regression diagnostics. *JASA*, 91:1659–1667, 1996.

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *JMLR*, 9:2015–2033, 2008.

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

C. M Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *JMLR*, 11:2973–3009, 2010.

G. Blanchard, C. Schäfer, and Y. Rozenholc. Oracle bounds and exact algorithm for dyadic classification trees. In *Proc. COLT*, pages 378–392. Springer, 2004.

L. Bottou and C-J. Lin. Support vector machine solvers. *Large scale kernel machines*, pages 301–320, 2007.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM-PROBAB STAT*, 9:323–375, 2005.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

S. Boucheron and M. Thomas. Concentration inequalities for order statistics. *Electron. Commun. Probab.*, 17:1–12, 2012.

S. Boucheron and M. Thomas. Tail index estimation, concentration and adaptivity. *EJS*, 9:2751–2792, 2015.

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004.

L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

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

L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A Mueller, O Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. 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.

B. Cadre. Kernel estimation of density level sets. *JMVA*, 97:999–1023, 2006.

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

V. Chernozhukov. Extremal quantile regression. *Ann. Stat.*, pages 806–839, 2005.

S. Clémençon and J. Jakubowicz. Scoring anomalies: a m-estimation formulation. In *Proc. AISTATS*, volume 13, pages 659–667, 2013.

S. Clémençon and S. Robbiano. Anomaly Ranking as Supervised Bipartite Ranking. In *Proc. ICML*, 2014.

S. Clémençon and N. Vayatis. Nonparametric estimation of the precision-recall curve. In *Proc. ICML*, pages 185–192, 2009a.

S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Trans Inf Theory*, 55:4316–4336, 2009b.

S. Clémençon and N. Vayatis. Overlaying classifiers: a practical approach to optimal scoring. *Constr Approx*, 32:619–648, 2010.

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

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 *AEROSP CONF PROC*, pages 1–11, 2008.

S. Coles, J. Bawa, L. Trenner, and P. Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

S. Coles and J.A 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. *JMVA*, 101:2103–2117, 2010.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *Ann. Stat.*, pages 2300–2312, 1997.

A. Daouia, L. Gardes, and S. Girard. On kernel smoothing for extremal quantile regression. *Bernoulli*, 19:2557–2589, 2013.

A. Daouia, L. Gardes, S. Girard, and A. Lekina. Kernel estimators of extreme level curves. *Test*, 20: 311–333, 2011.

J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proc. ICML*, 2006.

L. De Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

L. de Haan and S. Resnick. Second-order regular variation and rates of convergence in extreme-value theory. *Ann. Prob*, pages 97–124, 1996.

L. de Haan and S.I. Resnick. Limit theory for multivariate sample extremes. *Z WAHRSCHEINLICHKEIT*, 40:317–337, 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. Stat.*, 17:1833–1855, 12 1989.

F. Denis, R. Gilleron, and F. Letouzey. Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348:70–83, 2005.

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.

R. Díaz-Uriarte and S.A. De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 2006.

H. Drees and X. Huang. Best attainable rates of convergence for estimators of the stable tail dependence function. *JMVA*, 64:25–47, 1998.

M. C. du Plessis, G. Niu, and M. Sugiyama. Class-prior estimation for learning from positive and unlabeled data. In *Proc. ACML*, volume 45, 2015.

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. Stat.*, 34:1987–2014, 2006.

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

J. H. J. Einmahl, J. Li, and R. Y. Liu. Thresholding events of extreme in simultaneous monitoring of multiple risks. *JASA*, 104:982–992, 2009.

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

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

J. HJ Einmahl and D. M Mason. Generalized quantile processes. *Ann. Stat.*, 20:1062–1078, 1992.

P. Embrechts, L. de Haan, and X. Huang. Modelling multivariate extremes. *Extremes and integrated risk management*, pages 59–67, 2000.

E. Eskin. Anomaly detection over noisy data using learned probability distributions. In *Proc. ICML*, pages 255–262, 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*, 27:861–874, 2006.

H. Federer. *Geometric Measure Theory*. Springer, 1969.

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

P.A. Flach. The geometry of ROC space: understanding ML metrics through ROC isometrics. In *Proc. ICML*, 2003.

A.-L. Fougeres, L. De Haan, and C. Mercadier. Bias correction in multivariate extremes. *Ann. Stat.*, 43:903–934, 2015.

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

Y. Freund and R.E Schapire. Experiments with a new boosting algorithm. In *Proc. ICML*, volume 96, pages 148–156, 1996.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001.

L. Gardes and S. Girard. A moving window approach for nonparametric estimation of the conditional tail index. *JMVA*, 99:2368–2388, 2008.

L. Gardes, S. Girard, and A. Lekina. Functional nonparametric estimation of conditional extreme quantiles. *JMVA*, 101:419–433, 2010.

R. Genuer, J.-M. Poggi, and C. Tuleau. Random forests: some methodological insights. *arXiv:0811.3619*, 2008.

R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recog. Letters*, 31:2225–2236, 2010.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63:3–42, 2006.

C. Gini. Variabilita e mutabilita. *Memorie di metodologia statistica*, 1912.

S. Girard. A hill type estimator of the weibull tail-coefficient. *Communications in Statistics-Theory and Methods*, 33:205–234, 2004.

S. Girard and P. Jacob. Frontier estimation via kernel regression on high power-transformed data. *JMVA*, 99:403–420, 2008.

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. Submitted to AISTATS, 2016a.

N. Goix, A. Sabourin, and S. Clémençon. Sparse Representation of Multivariate Extremes. NIPS 2015 Workshop on Nonparametric Methods for Large Scale Representation Learning, 2015a.

N. Goix, A. Sabourin, and S. Clémençon. Sparse Representation of Multivariate Extremes with Applications to Anomaly Detection. In the reviewing process of JMVA, July 2016b.

N. Goix, A. Sabourin, and S. Clémençon. Learning the dependence structure of rare events: a non-asymptotic study. In *Proc. COLT*, 2015b.

N. Goix, A. Sabourin, and S. Clémençon. On Anomaly Ranking and Excess-Mass Curves. In *Proc. AISTATS*, 2015c.

N. Goix, A. Sabourin, and S. Clémençon. Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. In *Proc. AISTATS*, 2016c.

N. Goix and A. Thomas. How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms? To be submitted, 2016.

J.A. Hartigan. Estimation of a convex density contour in two dimensions. *JASA*, 82:267–270, 1987.

B. M. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Stat.*, 3: 1163–1174, 1975.

T.K. Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20:832–844, 1998.

V.J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artif. Intell. Review*, 22:85–126, 2004.

X. Huang. Statistics of bivariate extreme values. *PhD thesis*, 1992.

S. Janson. On concentration of probability. *Contemporary combinatorics*, 11, 2002.

E. Jones, T. Oliphant, P. 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. *Ann. Stat.*, 25:435–477, 1997.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization (with discussion). *Ann. Stat.*, 34:2593–2706, 2006.

M R. Leadbetter, G. Lindgren, and H. 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 *ICPR*, 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 *RAID*, pages 162–182. Springer, 2000.

B. Liu, W. S. Lee, P. S Yu, and X. Li. Partially supervised classification of text documents. In *Proc. ICML*, volume 2, pages 387–394, 2002.

F.T. Liu, K.M. Ting, and Z-H. Zhou. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data*, 6:3, 2012.

F.T. Liu, K.M. Ting, and Z.H. Zhou. Isolation Forest. In *ICDM*, pages 413–422, 2008.

G. Louppe. Understanding random forests: From theory to practice. *arXiv:1407.7502*, 2014. PhD Thesis.

M. Markou and S. Singh. Novelty detection: a review part 1: statistical approaches. *Signal proc.*, 2003.

D. M. Mason and W. Polonik. Asymptotic normality of plug-in level set estimates. *The Annals of Applied Probability*, 19:1108–1142, 2009.

P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse*, 9: 245–303, 2000.

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIV, volume 1896 of Lecture Notes in Mathematics.* Springer-Verlag, 2007.

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.

F. Mordelet and J-P Vert. A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

D.W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *JASA*, 86:738–746, 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. *COMPUT NETW*, 51:3448–3470, 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*, 12:2825–2830, 2011.

W. Polonik. Measuring Mass Concentrations and Estimating Density Contour Cluster-An excess Mass Approach. *Ann. Stat.*, 23:855–881, 1995.

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

W. Polonik. The silhouette, concentration functions and ml-density estimation under order restrictions. *Ann. Stat.*, 26:1857–1877, 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 *Proc. ICML*, volume 98, pages 445–453, 1998.

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

J.A Quinn and M. Sugiyama. A least-squares approach to anomaly detection in static and sequential data. *Pattern Recognition Letters*, 40:36–40, 2014.

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

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

P. Rigollet and R. Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15:1154–1178, 2009.

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

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

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

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

E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel. On Evaluation of Outlier Rankings and Outlier Scores. In *SDM*, pages 1047–1058. SIAM, 2012.

C. Scott and G. Blanchard. Novelty detection: Unlabeled data definitely help. In *Proc. AISTATS*, pages 464–471, 2009.

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

J. Segers. Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, 18:764–782, 08 2012a.

J. Segers. Max-stable models for multivariate extremes. *REVSTAT - Statistical Journal*, 10:61–82, 2012b.

C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE MC2R*, 5:3–55, 2001.

J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

T. Shi and S. Horvath. Unsupervised learning with random forest predictors. *J. Comp. Graph. Stat.*, 15, 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. Stat.*, 15:1174–1207, 09 1987.

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

A. J. Smola, L. Song, and C. H. Teo. Relative novelty detection. In *Proc. AISTATS*, volume 12, pages 536–543, 2009.

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

A. Stephenson. Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6:49–59, 2003.

A.G. Stephenson. High-dimensional parametric modelling of multivariate extreme events. *Australian & New Zealand Journal of Statistics*, 51:77–88, 2009.

M. Sugiyama. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*, 93:2690–2701, 2010.

V. Svetnik, A. Liaw, C. Tong, J C. Culberson, R.P Sheridan, and B.P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *J. Chem. Inf. Model.*, 43:1947–1958, 2003.

M. Tavallaee, E. Bagheri, W. Lu, and A.A. Ghorbani. A detailed analysis of the kdd cup 99 data set. In *IEEE CISDA*, volume 5, pages 53–58, 2009.

JA Tawn. Modelling multivariate extreme value distributions. *Biometrika*, 77:245–253, 1990.

D.MJ Tax and R.PW Duin. Uniform object generation for optimizing one-class classifiers. *JMLR*, 2: 155–173, 2002.

A. Thomas, V. Feuillard, and A. Gramfort. Calibration of One-Class SVM for MV set estimation. In *DSAA*, pages 1–9, 2015.

M. Thomas. *Concentration results on extreme value theory*. PhD thesis, Université Paris Diderot Paris 7, 2015.

A. B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Stat.*, 25:948–969, 1997.

S. Van Der Walt, S C Colbert, and G. Varoquaux. The numpy array: a structure for efficient numerical computation. *Comput Sci Eng*, 13:22–30, 2011.

V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *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.

R. Vert. *Theoretical insights on density level set estimation, application to anomaly detection*. PhD thesis, Paris 11, 2006.

K. Viswanathan, L. Choudur, V. Talwar, C. Wang, G. Macdonald, and W. Satterfield. Ranking anomalies in data centers. In *2012 IEEE Network Operations and Management Symposium*, pages 79–87, 2012.

JA. Wellner. Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Z WAHRSCHEINLICHKEIT*, 45:73–88, 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*, volume 8, pages 275–300, 2000.