Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere

Albert Thomas
LTCI, Télécom ParisTech
Université Paris-Saclay
Airbus Group Innovations, Suresnes

Stephan Clémençon
LTCI, Télécom ParisTech
Université Paris-Saclay

Alexandre Gramfort
LTCI, Télécom ParisTech
Université Paris-Saclay

Anne Sabourin
LTCI, Télécom ParisTech
Université Paris-Saclay

Abstract

Extreme regions in the feature space are of particular concern for anomaly detection: anomalies are likely to be located in the tails, whereas data scarcity in such regions makes it difficult to distinguish between large normal instances and anomalies. This paper presents an unsupervised algorithm for anomaly detection in extreme regions. We propose a Minimum Volume set (MV-set) approach relying on multivariate extreme value theory. This framework includes a canonical pre-processing step, which addresses the issue of output sensitivity to standardization choices. The resulting data representation on the sphere highlights the dependence structure of the extremal observations. Anomaly detection is then cast as a MV-set estimation problem on the sphere, where volume is measured by the spherical measure and mass refers to the angular measure. An anomaly then corresponds to an unusual observation given that one of its variables is large. A preliminary rate bound analysis is carried out for the learning method we introduce and its computational advantages are discussed and illustrated by numerical experiments.

1 Introduction

Motivated by a wide variety of applications including fraud detection, monitoring of complex networks and aviation safety management, unsupervised anomaly detection has recently received much attention in the machine learning literature. In many situations, measurements are considered as abnormal when they are located far from central measures such as the sample mean, rarity somehow replacing labeling from this perspective. This view has been extensively considered in the one-dimensional setup and has lead to a variety of statistical techniques for anomaly detection based on parametric representation of the tail of the observed univariate probability distribution, relying on *extreme value theory* (EVT) (see *e.g.* [3, 11, 17] among others). When a complex system is monitored by several physical variables, raising an alert at each extreme value of one of its physical variables can lead to high false alarm rates. A way to reduce this false alarm rate is to study the multivariate distribution of the set of observations such that at least one of their variables is large.

The purpose of the present paper is to promote an anomaly detection algorithm for such multivariate problems, based on *multivariate* EVT. In the suggested framework, *extreme* data are observed values $\boldsymbol{X}$ such that $\|\boldsymbol{X}\|_\infty$ is large, denoting by $\|\cdot\|_\infty$ the sup norm on $\mathbb{R}^d$. *Anomalies* among extremes are those which *direction* $\boldsymbol{X}/\|\boldsymbol{X}\|_\infty$ is unusual, which is an appropriate model for anomalies in many applications. We emphasize that from this perspective, some extreme data may be normal (not abnormal). The purpose of this paper is precisely to detect anomalies among these extremes. Note that there may well be non extreme data which are anomalies. These anomalies are outside our scope, since the suggested approach is only designed for extreme regions and may be combined with any state-of-the-art algorithm on non extreme regions, as detailed in [8]. The main idea consists in applying a classical multivariate anomaly detection approach, that is based on minimum volume sets (MV-sets in short) estimation, to a transformation of the original data emphasizing the dependence structure

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

of the *extreme* ones. Given a random vector $\boldsymbol{X}$ taking its values in $\mathcal{X} \subset \mathbb{R}^d$ with $d \geq 1$, MV-sets correspond to subsets of the feature space $\mathcal{X} \subset \mathbb{R}^d$ where the probability distribution F of the random variable $\boldsymbol{X}$ is most concentrated. More precisely, given a measure $\lambda(dx)$ of reference on the space $\mathcal{X}$ equipped with its Borel σ-algebra $\mathcal{B}(\mathcal{X})$ and $\alpha \in (0,1)$, a MV-set of level α for $\boldsymbol{X}$ is any solution Ω^*_α of the problem:

$$\min_{\Omega \in \mathcal{B}(\mathcal{X})} \lambda(\Omega) \text{ subject to } \mathbb{P}\{\boldsymbol{X} \in \Omega\} \geq \alpha\,. \qquad (1)$$

State-of-the-art methods for MV-sets estimation and anomaly detection (*e.g.* [19, 18, 13]) are usually sensitive to scaling effects and consider a fixed level $\alpha \in (0,1)$ in their theoretical analysis (*e.g.* [19, 23]) whereas the approach we suggest is concerned with extreme regions (the level α tends to 1) and is insensitive to scaling effects. The present work is related to [8, 9], as we also apply multivariate EVT to unsupervised anomaly detection. Yet, their approach is based on dimensionality reduction, which amounts to identifying the support of the distribution of the *directions* of extremes (the *angular measure*, see Section 2.2). Here the goal is different: we focus on moderate dimensional problems (typically, small subgroups of features obtained after a dimensionality reduction step) and the task is to detect anomalies in this moderate dimensional space. To do so we estimate MV-sets of the angular measure, contrary to [8, 9] who only identify a support.

The paper is organized as follows. In section 2, basic notions of the theory of MV-sets are briefly recalled, together with related statistical results. The multivariate EVT framework considered throughout the paper and its properties are also detailed. Section 3 explains the generic approach we propose for anomaly detection on extreme regions, based on a rank transformation of the data followed by a pseudo-polar decomposition. The construction of *critical regions* mainly relies on MV-set estimation methods on the sphere, yielding a convenient representation of the most probable *directions* of extreme events. The main theoretical result of the paper is a statistical guarantee concerning empirical recovery of MV-sets on the sphere. Numerical results are shown in Section 4. An extended analysis, tackling model selection issues in particular can be found in the Supplementary Material, together with detailed technical proofs.

2 Background and Preliminaries

As a first go, we recall key concepts pertaining to the theory of MV-sets, as well as known results related to their statistical recovery. Next, we describe the statistical setting we consider here for anomaly detection, relying on the multivariate heavy-tail assumption, and some key related properties which will be involved in the subsequent analysis. The indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$ and the Dirac mass at any point a by δ_a. Throughout the paper, vectors are denoted by bold letters, uppercase when random and lowercase otherwise. Finally, for any $(c, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^d$, we set $c\boldsymbol{x} = (cx^{(1)},\ \ldots,\ cx^{(d)})$, $[0, \boldsymbol{x}] = [0, x^{(1)}] \times \ldots \times [0, x^{(d)}]$ and $\boldsymbol{c}$ means the vector $(c,\ \ldots,\ c) \in \mathbb{R}^d$. We denote by $\mathbb{S}_{d-1}$ the intersection of the unit sphere $\{\boldsymbol{x} \in \mathbb{R}^d :\ \|\boldsymbol{x}\|_\infty = 1\}$ with the positive orthant $\mathbb{R}^d_+$. $\mathbb{S}_{d-1}$ is thus the positive orthant of the unit hypercube.

2.1 MV-sets and Anomaly Detection

The concept of Minimum Volume set (MV-set) has been introduced for the purpose of defining the regions where a multivariate random variable $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(d)})$ takes its values with highest/smallest probability, generalizing the well-known notion of quantile for 1-dimensional distributions, see *e.g.* [5, 15]. Denote by $\mathcal{X} \subset \mathbb{R}^d$ the space where the r.v. $\boldsymbol{X}$ takes its values and let $\alpha \in (0,1)$ and λ be a σ-finite measure of reference on $\mathcal{X}$ equipped with its Borel σ-algebra $\mathcal{B}(\mathcal{X})$, any solution of the minimization problem (1) is called a MV-set of level α. Throughout the paper, we assume that $\boldsymbol{X}$'s distribution F is absolutely continuous w.r.t. λ and denote by $f(\boldsymbol{x}) = dF/d\lambda(\boldsymbol{x})$ the related density. For any $\alpha \in (0,1)$, under the assumption that the density f is bounded and $f(\boldsymbol{X})$ has a continuous distribution F_f, one may show [15] that the set $\Omega^*_\alpha = \{\boldsymbol{x} \in \mathcal{X} :\ f(\boldsymbol{x}) \geq F_f^{-1}(1-\alpha)\}$ is the unique solution of the *minimum volume set* problem (1), where the generalized inverse function of any cumulative distribution function (*c.d.f.*) $K(t)$ on $\mathbb{R}$ is denoted by $K^{-1}(u) = \inf\{t \in \mathbb{R} :\ K(t) \geq u\}$. For high values of the mass level α, minimum volume sets are expected to contain the modes of the distribution, whereas their complementary sets correspond to *abnormal observations*. Refer to [5, 15] for details on minimum volume set theory and to [19, 23] for related statistical learning results.

Empirical MV**-sets.** A mass level $\alpha \in (0,1)$ being preliminarily fixed, estimating an empirical MV-set consists in building from training data $\boldsymbol{X}_1,\ \ldots,\ \boldsymbol{X}_n$ an estimate of a specific density level set Ω^*_α by solving a natural statistical counterpart of problem (1):

$$\min_{\Omega \in \mathcal{G}} \lambda(\Omega) \text{ subject to } \widehat{F}_n(\Omega) \geq \alpha - \psi_n\,, \qquad (2)$$

where ψ_n plays the role of a *tolerance parameter*, and where optimization is restricted to a subset $\mathcal{G}$ of $\mathcal{B}(\mathcal{X})$. The class $\mathcal{G}$ is supposed to be rich enough to include Ω^*_α or a reasonable approximation of it. It is ideally made of sets Ω whose volume $\lambda(\Omega)$ can be efficiently computed or estimated, *e.g.* by Monte-Carlo simulation. The empirical distribution based on the training sample (or a smoothed version of the latter) $\widehat{F}_n = (1/n)\sum_{i=1}^n \delta_{\boldsymbol{X}_i}$ replaces F and the tolerance parameter Ψ_n is chosen of the same order of magnitude as the supremum $\sup_{\Omega \in \mathcal{G}} |\widehat{F}_n(\Omega) - F(\Omega)|$.

Under usual complexity assumptions on the class $\mathcal{G}$ combined with an appropriate choice of ψ_n, non-asymptotic statistical guarantees for solutions $\widehat{\Omega}_\alpha$ of (2) are given in [19], together with algorithmic approaches to compute such solutions.

2.2 Multivariate Extreme Value Theory

In many statistical problems, risks are empirically described by sample means (*i.e.* the empirical classification error in supervised learning) and the theoretical validity of inference/learning methods based on such statistics is established by investigating how they deviate from their expectations. In contrast, EVT is dedicated to studying phenomena ruled by exceptionally large observations rather than averaging effects. Its main purpose is to provide models for learning the unusual rather than the usual, in order to assess the probability of occurrence of rare events. In risk monitoring, a quantity of major interest in the univariate situation is the $(1-\alpha)$-quantile of the distribution F of a r.v. X, for a given probability $1-\alpha$, that is $x_\alpha = \inf\{x \in \mathbb{R},\ \mathbb{P}\{X > x\} \leq 1-\alpha\}$. Whereas for moderate values of α, the statistic $x_{\alpha,n} = \inf\{x \in \mathbb{R},\ 1/n\sum_{i=1}^{n}\mathbb{I}\{X_i > x\} \leq 1-\alpha\}$ provides a natural empirical estimation, the information carried by the finite sample $X_1,\ldots,X_n$ is not sufficient when α is very large and the statistic $x_{\alpha,n}$ becomes irrelevant. In this case, one may call on EVT for providing parametric estimates of x_α. EVT mainly boils down to modeling the distribution of the maxima (respectively the upper tail) as a Generalized Extreme Value (GEV) distribution, namely an element of the Gumbel, Fréchet or Weibull parametric families (respectively by a generalized Pareto distribution). The primal – and not too stringent – assumption is the existence of two sequences $\{a_n, n \geq 1\}$ and $\{b_n, n \geq 1\}$, the a_n's being positive, and a non-degenerate *c.d.f.* G such that

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

for all continuity points $x \in \mathbb{R}$ of G. If this assumption is fulfilled – it is the case for most textbook distributions – then the tail behavior of F is essentially characterized by G, which is proved to belong to the parametric family of GEV distributions. Estimation of the tail's shape, scale and location parameters has been studied in great detail, see *e.g.* [1, 10, 21].

The multivariate analogue of the above display concerns the convergence of the tail probabilities,

$$\begin{aligned}\lim_{n\to\infty} n\,\mathbb{P}\Bigg\{&\frac{X^{(1)}-b_n^{(1)}}{a_n^{(1)}} \geq x^{(1)} \text{ or } \ldots \text{ or}\\ &\qquad\frac{X^{(d)}-b_n^{(d)}}{a_n^{(d)}} \geq x^{(d)}\Bigg\} = -\log \boldsymbol{G}(\boldsymbol{x}),\end{aligned} \tag{3}$$

where $a_n^{(j)} > 0$ and $b_n^{(j)} \in \mathbb{R}$ are entirely determined by the margins F_j and $\boldsymbol{G}$ is a non-degenerate multivariate *c.d.f.*.

Standardization and angular measure. The latter convergence is hardly workable as it is, since the sequences $a_n^{(j)}, b_n^{(j)}$ are unknown and $\boldsymbol{G}$ has no finite-dimensional parametrization nor generic structure. A customary first step consists in applying a specific increasing transform $T : \mathbb{R}^d \to \mathbb{R}^d$ such as $T(\boldsymbol{X}) = \boldsymbol{V}$ where for all $j \in \{1,\ldots,d\}$, $V^{(j)} = 1/(1-F_j(X^{(j)}))$. This allows to work with marginally standardized variables and to eliminate the unknown normalizing constants. Indeed the random vectors $\boldsymbol{X}$ and $\boldsymbol{V} = (V^{(1)},\ldots,V^{(d)})$ share the same copula, and thus the same dependence structure. Also, as shown by Proposition 5.10 in [16], the multivariate tail convergence assumption (3) then boils down to suppose that marginal tail convergence occurs, and that the standardized vector $\boldsymbol{V}$ has a regularly varying tail, *i.e.* there exists a limit measure μ on the starred positive orthant $\boldsymbol{E} = [0,\infty]^d \setminus \{0\}$, such that

$$n\,\mathbb{P}\left\{\frac{V^{(1)}}{n} \geq v^{(1)} \text{ or } \cdots \text{ or } \frac{V^{(d)}}{n} \geq v^{(d)}\right\} \xrightarrow[n\to\infty]{} \mu[0,\boldsymbol{v}]^c \tag{4}$$

for all $v^{(j)} > 0,\ 1 \leq j \leq d$. The measure μ is known as the *exponent measure* and it has the homogeneity property: $\mu(t\,\cdot\,) = t^{-1}\mu(\,\cdot\,)$. In terms of polar coordinates $(r(\boldsymbol{v}),\theta(\boldsymbol{v})) = (\|\boldsymbol{v}\|_\infty, (1/\|\boldsymbol{v}\|_\infty)\boldsymbol{v})$ the homogeneity property permits to write the limit distribution as a tensor product: for any measurable $\Omega \subset \mathbb{S}_{d-1}$, $t > 1$,

$$\mu(\boldsymbol{v} : r(\boldsymbol{v}) > t, \theta(\boldsymbol{v}) \in \Omega) = \frac{1}{t}\Phi(\Omega)\,, \tag{5}$$

where the *angular measure* $\Phi(\Omega) = \mu(\boldsymbol{v} : r(\boldsymbol{v}) > 1, \theta(\boldsymbol{v}) \in \Omega)$ can be turned into a probability distribution as $\Phi(\mathbb{S}_{d-1}) < \infty$. The measure Φ thus describes the probability distribution of the directions formed by the most extreme realizations. Notice incidentally that no parametric representation for Φ is provided by the theory. Any finite measure on $\mathbb{S}_{d-1}$ is a possible angular measure. The choice of the sup norm is mathematically convenient for the analysis of the error (Section 3.2), but alternative choices (such as any $\|\cdot\|_p$ norm, $p \geq 1$) would also be valid in a multivariate EVT setting.

Anomalies and extremal dependence structure. Our novel anomaly detection approach relies on a multivariate *Peaks-Over-Threshold* analysis. The focus is on the dependence structure of the components $X^{(j)}$ of large observations $\boldsymbol{X}$, where *large* means that at least one of the $X^{(j)}$'s is large: *e.g.* can some variables be large simultaneously or can only one variable be large at a time? In this context, given that at least one feature $X^{(j)}$ is large, $\boldsymbol{X}$ is an anomaly if it deviates from a characterization of the dependence structure of such observations (those $\boldsymbol{X}$'s such that at least one of the $X^{(j)}$ is large).

The starting point is the combination of (4) and (5) which imply that for $\Omega \subset \mathbb{S}_{d-1}$,

$$\mathbb{P}\big(r(\boldsymbol{V}) > t,\ \theta(\boldsymbol{V}) \in \Omega\big) \underset{t\to\infty}{\sim} \frac{1}{t}\Phi(\Omega)\,. \tag{6}$$

The angular measure Φ thus encapsulates the structure of the tail. Figure 1 illustrates this fact by showing extreme samples (in black) projected on the sup norm sphere. Two samples are generated from two distinct bi-dimensional extreme value logistic distributions (refer to Section 4.1 for the model description). The first (respectively, second) sample has a high (respectively, small) coefficient of dependence. Figure 1(a) shows that in the strongly dependent case, angular data are mostly concentrated around the angle $\pi/4$ whereas in the weakly dependent case (Figure 1(b)), angular data lie mostly around the two axes. In the limiting cases of these two situations, the angular measure would degenerate respectively into a single Dirac mass located at the angle $\pi/4$ and two Dirac masses at angles 0 and $\pi/2$.

Recovering MV-sets of high mass (*i.e.* corresponding to high values of α) for the angular measure Φ gives access to the most probable directions of extremes. In the case where the angular component alone should be considered for anomaly detection, those angular MV-sets would allow to pin the complementary sets as abnormal.

Remark 1 (Anomaly score)**.** *In practice, the radial part does play a role (see equation* (6)*) and we shall define an anomaly score which is a product of a radial score and an angular score based on a family of nested* MV*-sets (see Section 4).*

It is noteworthy to mention that the choice of standardized variables $\boldsymbol{V}$ fully avoids scaling effects due to unit choices and appears as well-founded in a variety of practical situations.

Motivated by these preliminary observations and in order to build critical regions for anomaly detection, the issue of estimating MV-sets of a sub-asymptotic version of the angular probability distribution is considered in the next section, from a theoretical and practical perspective.

3 MV-set Estimation on the Sphere

We now rigorously formulate the MV-sets statistical problem on the sphere. Denoting by $\mathcal{B}(\mathbb{S}_{d-1})$ the Borel σ-algebra on the sphere, the generic goal in a MV-set context is to recover from training observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ which are independent copies of the generic heavy-tailed r.v. $\boldsymbol{X}$, a solution of the problem $\min_{\Omega\in\mathcal{B}(\mathbb{S}_{d-1})} \lambda_d(\Omega)$ subject to $\Phi(\Omega) \geq \alpha$. Note that $\alpha \in (0, \Phi(\mathbb{S}_{d-1}))$, instead of $\alpha \in (0,1)$, as Φ is not a probability distribution.

In practice, the angular measure Φ is an asymptotic object, whereas the data at hand is non asymptotic. Also, it may be argued from a practical perspective that our interest lies in large, but non asymptotic regions $\{\boldsymbol{x} : r(T(\boldsymbol{x})) > t\}$. Consider thus the *sub-asymptotic* angular measure at finite level t, $\Phi_t(\Omega) = t\,\mathbb{P}(r(\boldsymbol{V}) > t, \theta(\boldsymbol{V}) \in \Omega)$ and notice from (6) that $\Phi_t(\Omega) \to \Phi(\Omega)$ as $t \to \infty$. In the sequel we shall thus consider the modified, non asymptotic optimization problem

$$\min_{\Omega\in\mathcal{B}(\mathbb{S}_{d-1})} \lambda_d(\Omega) \quad \text{subject to} \quad \Phi_t(\Omega) \geq \alpha\,. \tag{7}$$

In order to ensure existence and uniqueness of the solution of this optimization problem, we consider the following assumptions, which are commonly used in the MV-set literature to ensure the existence and uniqueness of the MV-set optimization problem [15].

- $\boldsymbol{A_1}$ For any $t > 1$, the distribution $\Phi_t(\,\cdot\,)$ is absolutely continuous w.r.t. the Lebesgue measure λ_d on $\mathbb{S}_{d-1}$ with density ϕ_t. In addition, the r.v. $\phi_t(\theta(\boldsymbol{V}))$ has no flat parts: $\forall c > 0$, $\mathbb{P}\{\phi_t(\theta(\boldsymbol{V})) = c\} = 0$.
- $\boldsymbol{A_2}$ The density $\phi_t(\theta)$ of $\Phi_t(\,\cdot\,)$ is uniformly bounded : $\sup_{t>1,\theta\in\mathbb{S}_{d-1}} \phi_t(\theta) < \infty$.

Given assumptions $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ one can show that (7) has a unique solution, given by the density level set $B^*_{\alpha,t} = \{\boldsymbol{\theta} \in \mathbb{S}_{d-1} :\ \phi_t(\boldsymbol{\theta}) \geq K^{-1}_{\Phi_t}(\Phi(\mathbb{S}_{d-1}) - \alpha)\}$, where $K_{\Phi_t}(y) = \Phi_t(\{\boldsymbol{\theta} \in \mathbb{S}_{d-1} :\ \phi_t(\boldsymbol{\theta}) \leq y\})$.

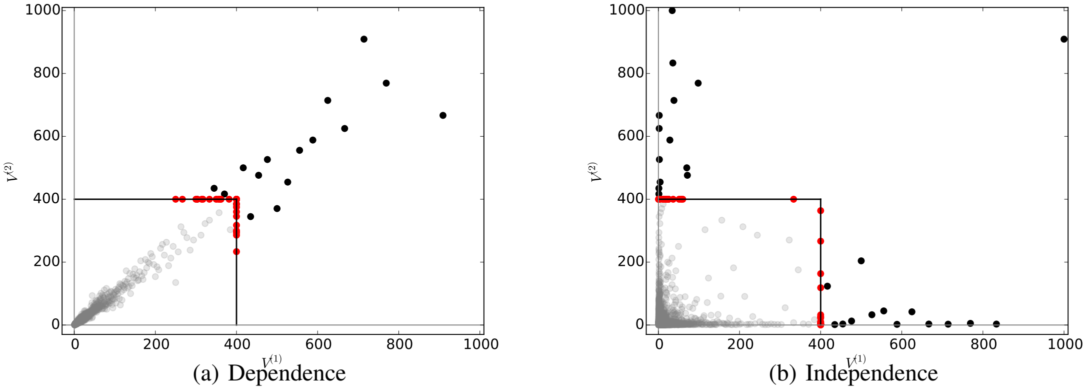
The general method described and studied in the next section consists in replacing in (7) the angular measure Φ_t by a sharp estimate, involving a fraction of the original observations (*i.e.* the most extreme observations).

3.1 Empirical MV-sets on the Sphere

Additional notations are required. For any $j \in \{1, \ldots, d\}$, the j-th empirical marginal c.d.f. is denoted by $\widehat{F}_j(u) = (1/n)\sum_{i=1}^n \mathbb{I}\{X_i^{(j)} \leq u\}$, $u \in \mathbb{R}$. A natural empirical version of the $\boldsymbol{V}_i$'s is obtained by means of a rank transformation, $\boldsymbol{V}_i = \widehat{T}(\boldsymbol{X}_i)$, where $\widehat{T}$ is defined for all $i \in \{1, \ldots, n\}$ by

$$\widehat{T}(\boldsymbol{X}_i) \overset{def}{=} \left(\frac{1}{1-\widehat{F}_1(X_i^{(1)})}, \ldots, \frac{1}{1-\widehat{F}_d(X_i^{(d)})}\right)\,. \tag{8}$$

This data standardization is widely used in multivariate EVT to study dependence among extremes, see [1] and the references therein for instance. From a practical angle, it is of disarming simplicity and fully avoids any distributional assumptions for the margins. Of course, the feature variables (8) are not independent anymore and analyzing the accuracy of an estimate of the angular distribution Φ involved in (5) based on the latter is far from straightforward. However, it has been shown in [4, 6] that using the rank transformed variables $\widehat{\boldsymbol{V}}_i$'s instead of the probability integral transformed ones $\boldsymbol{V}_i$ does not damage the asymptotic



(a) Dependence

(b) Independence

Figure 1: Illustration of the directions $\theta(\boldsymbol{V})$ obtained with a sample generated from a logistic model with a high coefficient of dependence (a) and a small coefficient of dependence (b). Non extreme samples are in gray, extreme samples in black and directions $\theta(\boldsymbol{V})$ (extreme samples projected on the sup norm sphere) in red. Note that not all extreme samples are shown as the plot has been truncated for a better visualization. However all projections on the sphere are shown.

properties of the empirical estimator of the angular measure (in dimension 2, under suitable regularity assumptions). In arbitrary dimension, Goix et al. [7] have obtained a similar result for the finite sample case, concerning an alternative characterization of the angular measure, which is an integrated version of Φ.

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

Statistical guarantees for the general algorithm 1 and a practical method for solving the optimisation problem (9) it involves are detailed in the following subsections. As shall be seen, from a practical perspective, a crucial advantage of the approach we promote lies in the compactness of the feature space $\mathbb{S}_{d-1}$ used to detect abnormal directions. Our analysis proceeds as if the marginal distributions were known, *i.e.* as if the true transformed variables $\boldsymbol{V}_i$'s were observables. Controlling the additional sample error induced by the discrepancy $\widehat{\boldsymbol{V}}_i - \boldsymbol{V}_i$ is reserved for future work.

3.2 Main Result

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

Algorithm 1 Empirical estimation of an angular MV-set

Inputs: Training data set $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, $k \in \{1, \ldots, n\}$, mass level α, tolerance $\psi_k(\delta)$, confidence level $1-\delta$, collection $\mathcal{G}$ of subsets of $\mathbb{S}_{d-1}$

Standardization: Apply the rank-transformation (8) to the $\boldsymbol{X}_i$'s, yielding the empirically marginally standardized vectors $\widehat{\boldsymbol{V}}_i = \widehat{T}(\boldsymbol{X}_i)$, $i = 1, \ldots, n$.

Thresholding: Retain the indexes

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

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

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

$$\min_{\Omega \in \mathcal{G}} \lambda_d(\Omega) \text{ subject to } \widehat{\Phi}_{n,k}(\Omega) \geq \alpha - \psi_k(\delta) \,. \tag{9}$$

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

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

Fix a mass level α and $\delta \in (0,1)$ and consider the empirical MV*-set $\widehat{\Omega}_\alpha$ solution of* (9) *related to the tolerance*

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

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

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

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

As expected, the rate of statistical recovery of the solution $B^*_{\alpha,n/k}$ of (7) when $t = n/k$ is of order $O_{\mathbb{P}}(\sqrt{1/k})$, the learning procedure involving the $|\mathcal{I}| \in [k, dk]$ most extreme standardized observations only. Before presenting a practical implementation of the approach analyzed here, a few remarks are in order.

Remark 2 (FINITENESS OF THE CLASS AND LEARNING RATE)**.** *The argument originally developed in [7] for controlling the accuracy of an empirical estimation of the stable tail dependence function (STDF) is crucially exploited to cope with the dependence structure of the transformed variables* (8). *The finite class assumption fits our purposes in the present paper, since we consider unions of rectangles paving the sphere (see Section 3.3). A minor modification of the argument in the supplementary material would allow to replace* $\log(|\mathcal{G}|)$ *with* $V_{\mathcal{G}} \log(dke/V_{\mathcal{G}})$*, where* $V_{\mathcal{G}}$ *is the VC-dimension of the class* $\mathcal{G}$*, and* dk *is an upper bound for the average number of points hitting the extreme regions (see the proof of Lemma 1 in the Supplementary Material). Then the learning rate bound given by the result above is of order* $O_{\mathbb{P}}(\sqrt{(\log k)/k})$*, as expected, since* $O(k)$ *observations are actually involved in the learning procedure, due to the thresholding stage.*

Remark 3. (ON THE CONTINUITY ASSUMPTION) *In order to place oneself in the framework of Theorem 1 (i.e. in the situation where the angular measure is absolutely continuous w.r.t. Lebesgue measure on the sphere), applying a preliminary dimension reduction technique to the original observations in the extremes can be necessary. It is the precisely the goal of the methods proposed in [8, 9] (see also [2]) to identify possible degenerate components of the angular measure, as well as subsets of variables forming random subvectors fulfilling Theorem 1's assumptions.*

3.3 Paving the Sphere

We now address the issue of solving (9) from a computational perspective. As a first go, we build empirical MV-sets on the sphere by binding together elementary subsets S of $\mathbb{S}_{d-1}$ with same volume (*i.e.* same Lebesgue measure $\lambda_d(S)$).

Again, empirical estimation $\widehat{\Phi}_{n,k}$ of the angular measure is based on the fraction $\{\boldsymbol{\theta}_i : \ i \in \mathcal{I}\}$ of the transformed data and we consider the partition of $\mathbb{S}_{d-1}$ in dJ^{d-1} hypercubes S_j with same volume as shown in Figure 2.

We therefore consider the class $\mathcal{G}$ that corresponds to the class $\mathcal{G}_J$ of subsets obtained as union of cubes S_j. In this case, $|\mathcal{G}| = \exp(dJ^{d-1}\log 2)$. Figure 2 shows an example of such a partition for $d = 3$ and $J = 5$. Sorting the elements by decreasing order with respect to the number of samples they contain and binding them together until reaching a mass greater than $\alpha - \psi_k(\delta)$ yields $\widehat{\Omega}_\alpha$ (see [19]).

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

The number of unique signatures is equal to the number of non empty hypercubes of the partition and the number of identical signatures is equal to the number of samples in the corresponding hypercube. We have therefore identified all the non empty hypercubes and the number of samples in each of them. Using Algorithm 2 we then obtain an estimated MV-set with level mass α, *i.e.*, the solution of (9).

Algorithm 2 Solution of (9) when $\mathcal{G}$ is the regular grid on $\mathbb{S}_{d-1}$

Sorting: Sort the elementary subsets S_j so that: $\widehat{\Phi}_{n,k}(S_{(1)}) \geq \ldots \geq \widehat{\Phi}_{n,k}(S_{(J)})$.

Concatenation: Bind together the elementary subsets sequentially, until the empirical angular measure of the resulting set exceeds $\alpha - \psi_k(\delta)$, yielding the region

$$\widehat{\Omega}_{J,\alpha} = \bigcup_{j=1}^{J(\alpha)} S_{(j)}\,, \qquad (10)$$

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

Remark 4. *While the complexity of the algorithm is linear in the dimension d, this approach suffers from the curse of dimensionality. Indeed, as the number of hypercubes increases exponentially with d, only a small proportion of hypercubes will be non-empty and the solution will tend to overfit.*

Remark 5. *When implementing the hash function we have to carefully deal with the samples* $\boldsymbol{\theta}$ *that are located on the edges of* $\mathbb{S}_{d-1}$, i.e.*, such that at least two of their components are equal to 1. Under assumption* $\boldsymbol{A}_1$*, the probability of a sample* $\boldsymbol{\theta}$ *to be located on an edge of* $\mathbb{S}_{d-1}$ *is equal to 0. However it is not always the case in practice, especially if we use the empirical marginals for the standardization* *step. The hash function defined above assigns a signature to an edge sample $\boldsymbol{\theta}$ that is equal to none of the signatures of the adjacent hypercubes of $\boldsymbol{\theta}$. Therefore we arbitrarily assign such samples to one of their adjacent hypercubes.*

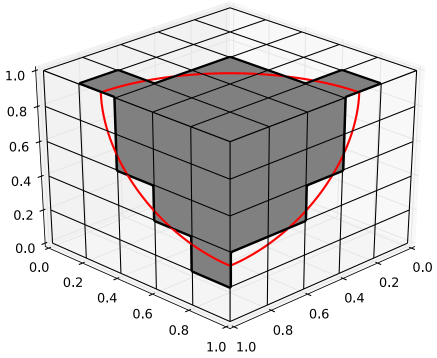


Figure 2: Estimated angular MV-set on the sphere based on Gaussian data. In red, the border of the true MV-set with mass at least 0.9. In gray the estimated MV-set with relative level mass 0.9 and $J = 5$.

Toy example. Figure 2 illustrates the MV-set obtained *via* algorithm 2 on a simple example. Here $d = 3$, so that $\mathbb{S}_2$ has 3 faces. Angular points on each faces are generated according to truncated bivariate Gaussian distributions on each face, centered at $(1, 1, 1)$ in the 3-dimensional space. The estimated MV-set shown in Figure 2 has a relative level mass of 0.9 and is obtained with $J = 5$.

Bias induced by the finite grid. Looking for the MV-set in the class $\mathcal{G}$ instead of all the measurable subsets of the sphere induces a bias which can be controlled with mild assumptions on the angular distribution, such as the box counting class introduced in [20] (see Supplementary Material).

Model selection. The resolution level J should be chosen with care as it can impact significantly the MV-set estimation procedure. This issue can be addressed through *complexity penalization* (see Supplementary Material). However for the numerical experiments we resort to cross validation selecting the resolution level giving an empirical angular mass close to α on a test set. Indeed if the grid is too coarse, the estimated MV-set should have an empirical measure much greater than α on a test set. Similarly, if the grid is too fine, the estimated MV-set will have an empirical measure much smaller than α on a test set.

3.4 Application to anomaly detection

As already mentioned in Remark 1, considering angular MV-sets only does not yield an optimal decision function, since the density of the largest observations includes a radial part. More precisely, in view of (6), the density (with respect to $dr \otimes d\theta$) on the most extreme regions is proportional to $\frac{1}{r^2}\phi(\theta)$. A standard approach in anomaly detection is to define a scoring function $\hat{s}$, which should be ideally proportional to the density, and then to declare as abnormal regions of the kind $\{x : \hat{s}(x) \leq s_0\}$, where s_0 can be tuned so that a given proportion of the samples are pinned as abnormal. It turns out that as a byproduct of our algorithm, we can also estimate a scoring function $\hat{s}_\theta$ on $\mathbb{S}_{d-1}$, such that the smaller $\hat{s}_\theta(\boldsymbol{\theta})$ is, the more abnormal the direction $\boldsymbol{\theta}$. We define $\hat{s}_\theta$ as the piecewise constant function defined on each hypercube of the partition of $\mathbb{S}_{d-1}$ by the number of samples it contains (see Figure 3(a)). One can then consider the scoring function on the whole space defined by

$$\hat{s}(r(\boldsymbol{V}), \theta(\boldsymbol{V})) = 1/r(\boldsymbol{V})^2 \cdot \hat{s}_\theta(\theta(\boldsymbol{V})). \qquad (11)$$

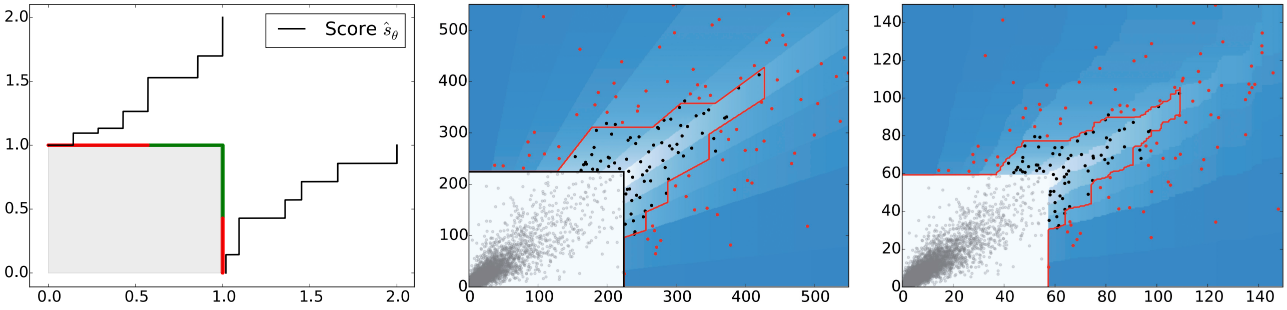
Again, the smaller $\hat{s}(r(\boldsymbol{V}), \theta(\boldsymbol{V}))$ is, the more abnormal $(r(\boldsymbol{V}), \theta(\boldsymbol{V}))$, i.e. $\boldsymbol{V}$. Using such a scoring function, observations with very large sup norm but with high angular score have a chance to be considered as anomalies, which would not be the case if the MV-set estimates on $\mathbb{S}_{d-1}$ only were considered.

4 Numerical Experiments

We first illustrate our approach on a bivariate simulated toy example. We then compare our approach to two state-of-the-art unsupervised anomaly detection algorithms, Isolation Forest [13] and One-Class SVM (OCSVM) [18], on five real data sets. We set $k = \sqrt{n}$ in all experiments. As we do not know the normalization constant $\Phi(\mathbb{S}_{d-1})$ of the angular measure, we use $|\mathcal{I}|$ to normalize the empirical angular measure and consider relative mass levels in $(0, 1)$. The penalty $\psi_k(\delta)$ in (9) would require k to be too large to allow us to consider it in practice. We therefore solve the optimization problem (9) setting $\psi_k(\delta) = 0$ (see Supplementary Material for the connection with the theoretical result). Finally, we use the implementation of Isolation Forest and OCSVM provided by Scikit-Learn [14].

4.1 Toy example - Logistic model

We consider a 2-dimensional sample $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ of size $n = 50000$ generated from an extreme value logistic model ([1], section 9.2.2) which is defined by its parametric c.d.f. $\boldsymbol{G}(\boldsymbol{x}) = \exp\Big\{ - \big(\sum_{j=1}^{d} (x^{(j)})^{-1/\beta}\big)^{\beta}\Big\}$, $x^{(j)} \geq 0$, for some parameter $\beta \in (0, 1]$. The smaller the parameter β, the more dependent the variables. We choose $\beta = 0.2$ (strong dependence). The data are simulated according to Algorithm 2.1 of [22]. We use the scoring function $\hat{s}$ defined by (11) and fix the anomaly threshold s_0 so that 30% of the extreme data are normal. Figure 3 displays the results: extremes outside the s_0 level set of $\hat{s}$ are considered



(a) Angular score on the sphere

(b) Standardized space $(V^{(1)}, V^{(2)})$

(c) Input space $(X^{(1)}, X^{(2)})$.

Figure 3: Illustration of our approach on a sample generated from a logistic model. Figure (a) shows the angular score obtained with our algorithm. In (b) and (c) the red contour shows the frontier between abnormal and normal regions. Non extreme samples are in gray and extreme anomalies are in red.

as anomalies.

4.2 Anomaly detection on real data sets

We now compare the performance of our approach (Algorithm 1, scoring function $\hat{s}$ (11)) with Isolation Forest and OCSVM on five classical anomaly detection data sets whose characteristics are summarized in Table 1. The shuttle, ann and forestcover data sets are available at the UCI Machine Learning repository [12]. For the shuttle data set, instances with label 4 are removed and instances with labels different than 1 are considered as abnormal. For the ann and forestcover data sets we only keep the continuous variables. The abnormal instances of the ann data set are those with label 1 or 2. For the forestcover data set, instances with label 2 are normal whereas instances with labels 4 and 5 are abnormal. The SF data set is obtained from the KDD Cup'99 intrusion detection data set following [24]. The anomaly class is the attack class. The http data set corresponds to all instances of the SF data set whose third feature is *http*. For the SF and http data sets only 10% of the whole data set is used.

Table 1: Data sets

Data set	n	d	**Anomaly ratio**
shuttle	85,849	9	7.2%
SF	699,691	4	0.3%
http	619,052	3	0.4%
ann	7,200	6	7.4%
forestcover	581,012	5	4.1%

In all experiments, the suggested algorithm, Isolation Forest and OCSVM are trained on half of the normal instances, chosen at random. The test set for both algorithms consists in all instances (normal and abnormal) not used in the training set. This test set is then restricted to the extreme region in accordance with the thresholding step of Algorithm 1 and performance is assessed with the available labels. For all data sets, the results are averaged over 10 experiments conducted with normal training samples selected at random.

Table 2: ROC-AUC

Data set	**OCSVM**	**Isolation Forest**	**Score $\hat{s}$**
shuttle	0.981	0.963	**0.987**
SF	0.478	0.251	**0.660**
http	**0.997**	0.662	0.964
ann	0.372	**0.610**	0.518
forestcover	0.540	0.516	**0.646**

Areas under the Receiver Operating Characteristic curve (ROC-AUC) obtained on all data sets are dispayed in Table 2. Our approach outperforms Isolation Forest and OCSVM in the extreme region on three out of five data sets and is never the worst one.

5 Conclusion

This paper addresses the issue of anomaly detection in extreme regions. The methodology we propose is based on statistical recovery of MV-sets for the angular measure on the sphere, a functional measure of the dependence structure of extreme observations. Anomalies correspond to unusual relative contributions of specific variables to the considered extreme event. Future work will aim at controlling the error induced by the empirical estimation of margins. Another natural extension will be to consider an adaptive paving of $\mathbb{S}_{d-1}$ instead of a fixed partition in elements of identical volume.

References

[1] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics, 2005.

[2] E. Chautru. Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics*, 9(1):383–418, 2015.

[3] D. A. Clifton, S. Hugueny, and L. Tarassenko. Novelty detection with multivariate extreme value statistics. *Journal of signal processing systems*, 65(3):371–389, 2011.

[4] J.H.J. Einmahl, L. De Haan, and V.I. Piterbarg. Nonparametric estimation of the spectral measure of an extreme value distribution. *Annals of Statistics*, pages 1401–1423, 2001.

[5] J.H.J. Einmahl and D.M. Mason. Generalized quantile process. *The Annals of Statistics*, 20:1062–1078, 1992.

[6] J.H.J. Einmahl and J. Segers. Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, pages 2953–2989, 2009.

[7] N. Goix, A. Sabourin, and S. Clémençon. Learning the dependence structure of rare events: a nonasymptotic study. In *Proceedings of the International Conference on Learning Theory, COLT'15*, 2015.

[8] N. Goix, A. Sabourin, and S. Clémençon. Sparse representation of multivariate extremes with applications to anomaly ranking. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS'16*, 2016.

[9] N. Goix, A. Sabourin, and S. Clémençon. Sparsity in multivariate extremes with applications to anomaly detection. *http://arxiv.org/abs/1507.05899*, 2016.

[10] B. M. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3(5):1163–1174, 09 1975.

[11] H.J. Lee and S.J. Roberts. On-line novelty detection using the kalman filter and extreme value theory. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4, 2008.

[12] M. Lichman. UCI machine learning repository, 2013.

[13] F.T. Liu, K.M. Ting, and Z-H. Zhou. Isolation forest. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, pages 413–422, 2008.

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

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

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

[17] S.J. Roberts. Extreme value statistics for novelty detection in biomedical signal processing. In *Advances in Medical Signal and Information Processing, 2000. First International Conference on (IEE Conf. Publ. No. 476)*, pages 166–172, 2000.

[18] B. Schölkopf, J. Platt, A. J. Shawe-Taylor, J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7), 2001.

[19] C. Scott and R. Nowak. Learning Minimum Volume Sets. *Journal of Machine Learning Research*, 7:665–704, 2006.

[20] C. Scott and R. D. Nowak. Minimax-Optimal Classification With Dyadic Decision Trees. *Information Theory, IEEE Transactions on*, 52(4):1335–1353, 2006.

[21] R. L. Smith. Estimating tails of probability distributions. *Ann. Statist.*, 15(3):1174–1207, 09 1987.

[22] A. Stephenson. Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6(1):49–59, 2003.

[23] R. Vert and J.P. Vert. Consistency and convergence rates of one-class SVM and related algorithms. *J. Machine Learning Research*, 17:817–854, 2006.

[24] K. Yamanishi, J-I. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Min. Knowl. Discov.*, 8(3):275–300, 2004.