Report: Sparse representation of multivariate extremes with applications to anomaly detection

Mathis Hardion

December 2023

I. Introduction

This report discusses the results, limits and implications of [1]. The latter deals with capturing dependency structures within multivariate extreme distributions in high dimensional settings, by expliciting through the angular measure a smaller subset of directions along which extremes may happen. If such directions can be expressed with a small number of coordinates, one can scale up algorithms suited to low dimensions to more complex problems. An estimator of the repartition of the angular measure's mass on sub-cones is constructed and VC-type non-asymptotic convergence bounds are derived. A numerical study is made to provide evidence that "sparse" dependencies occur in real datasets. The method is applied to detecting anomalies as extremes in uncommon directions in order to illustrate its relevance. This report is organized as follows. Section II introduces the notations and states the problem of interest in the paper, and critically addresses the assumptions made by the authors. Section III explains the main theoretical contributions of the paper and discusses their limits and implications. Section IV covers the application to anomaly detection while section V challenges a reimplementation of the paper's algorithm against other methods and further investigates parameter influence. Finally, section VI summarizes the main points of the discussion and gives tentative directions in which the paper could be expanded upon.

II. Problem formulation and hypotheses

The problem of interest in this paper can be summarized as finding the few main directions in which multivariate data may be extreme, so that fewer dimensions are enough to characterize its tail distribution. This is formalized as follows: let $\mathbf{X} = (X^1, \dots, X^d)$ be a random vector valued in $\mathbb{R}^d$ endowed with the infinity or supremum norm, denote $(F_j)_{1\leq j\leq d}$ the marginal cumulative distribution functions, $\mathbf{V} \coloneqq \left(\frac{1}{1-F_j(X^j)}\right)_{1\leq j\leq d}$. Consider the usual multivariate regular variation hypothesis, i.e. assume that we have a measure μ on $[\mathbf{0}, \boldsymbol{\infty}]^d \setminus \{\mathbf{0}\}$ such that

$$\forall \mathbf{v}, n\mathbb{P}\left(\frac{1}{n}V \in [\mathbf{0}, \mathbf{v}]^c\right) \underset{n\to\infty}{\longrightarrow} \mu\left([\mathbf{0}, \mathbf{v}]^c\right). \tag{1}$$

A "direction" is going to correspond to a group of nonzero coordinates, i.e. a subset of indices $\alpha \subset \{1, \dots, d\}$. The set of vectors along such a direction is denoted

$$\mathcal{C}_\alpha \coloneqq \left\{\mathbf{v} \geq 0 \;\middle|\; \|\mathbf{v}\|_\infty \geq 1, \mathbf{v}^{|\alpha} > 0, \mathbf{v}^{|\alpha^c} = 0\right\},$$

where for $\mathbf{v} = (v^1, \dots, v^d)$, $\mathbf{v}^{|\alpha}$ refers to the vector $\left(v^j\right)_{j\in\alpha}$, and α^c to the complementary set of α in $\{1, \dots, d\}$. The corresponding directions (i.e. unit vectors) are denoted $\Omega_\alpha \coloneqq S_\infty^{d-1} \cap \mathcal{C}_\alpha$ where S_∞^{d-1} is the sphere for the infinity norm. Considering the $\mathcal{C}_\alpha$s have boundary with possibly nonzero mass (they have empty interior, meaning their boundary is their closure) and thus the type of convergence in (1) does not necessarily hold for such sets, this motivates the introduction of truncated rectangles, namely for $\varepsilon > 0$,

$$R_\alpha^\varepsilon \coloneqq \left\{\mathbf{v} \geq 0 \;\middle|\; \|\mathbf{v}\|_\infty \geq 1, \mathbf{v}^{|\alpha} \geq \varepsilon, \mathbf{v}^{|\alpha^c} < \varepsilon\right\}.$$

That way, one can work with continuity sets of μ and thus obtain statistical estimates with convergence guarantees: one can show by splitting R_α^ε as the difference of two sets decreasing in ε and using the continuity from above property of measures that

$$\mu\left(R_\alpha^\varepsilon\right) \underset{\varepsilon\to 0}{\longrightarrow} \mu\left(\mathcal{C}_\alpha\right). \tag{2}$$

The problem is then, given i.i.d. samples of X's distribution, to build an estimator of $\mathcal{M} \coloneqq (\mu(\mathcal{C}_\alpha))_{\alpha \subset \{1,\dots,d\}}$ and to derive convergence bounds. In the case where the number of αs for which the above quantities are nonnegligible is low with respect to 2^d, one obtains a 'sparse' representation, and if such directions α contain only few coordinates, the representation is low-dimensional.

In order to give an answer to this problem, the authors make 3 major assumptions (in addition to the regular variation hypothesis) which we now discuss.

Assumption 1. The marginal cdfs $(F_j)_{1 \leq j \leq d}$ are continuous.

While this is a standard assumption in our context, it is not always verified, as it does not apply as soon as certain singletons have nonzero mass under one of the marginal distributions. The obvious counterexample is that of a variable valued in a discrete space, e.g. binary variables or integer measurements. A not so immediately apparent yet real example is that of variables valued in $\mathbb{R}_+$ taking the value 0 with nonzero probability, for instance power consumption of certain components of a system that are not required to run at all times. Therefore, this assumption is not as trivial as it may seem at first glance.

Assumption 2. For $\varnothing \neq \alpha \coloneqq \{i_1, \dots, i_r\} \subset \{1, \dots, d\}$, $\mu_\alpha(\cdot) \coloneqq \mu(\cdot \cap \mathcal{C}_\alpha)$ is absolutely continuous with respect to $dx_\alpha \coloneqq dx_{i_1} \dots dx_{i_r}$.

While assumption 2 implies assumption 1, they are not equivalent and there can be possible yet nonobvious cases where the latter applies yet not the former. Let us consider the case where one of the considered marginals is deterministically determined by another, i.e. $X^2 = f(X^1)$ for some increasing bijection f. Note that if one does not carefully examine the interpretation of each feature, such a dependancy might be difficult to notice by looking only at the data itself. Then, we have $F_2(x) = F_1(f^{-1}(x))$ and thus

$$V^2 = \frac{1}{1 - F_2(X_2)} = \frac{1}{1 - F_1(f^{-1}(X_2)} = V^1.$$

Consider $A_\varepsilon \coloneqq \{(v_1, v_2) \geq 1 | \, |v_1 - v_2| \leq \varepsilon\}$, so that

$$\begin{aligned} n\mathbb{P}\left(\frac{\mathbf{V}}{n} \in A_\varepsilon\right) &= n\mathbb{P}\left(V^1 \geq n, \left|V^1 - V^1\right| \leq \varepsilon\right) \\ &= n\mathbb{P}\left(V^1 \geq n\right) \\ &\underset{n \to \infty}{\longrightarrow} 1 \end{aligned}$$

by the regular variation hypothesis. For $\varepsilon_\ell = 2^{-\ell}$, the sequence $(A_{\varepsilon_\ell})_\ell$ is decreasing, meaning that for $A \coloneqq \bigcap_{\ell=0}^{+\infty} A_{\varepsilon_\ell}$,

$$\mu(A) = \lim_{\ell \to +\infty} \mu(A_{\varepsilon_\ell}) = 1$$

since μ is a measure. However, $A = \{(v_1, v_2) \geq 1 | v_1 = v_2\}$ has Lebesgue measure 0, meaning assumption 2 does not hold. This shows the importance of preprocessing and removing features adding no information.

Under assumption 2, and denoting $\Phi(\cdot) \coloneqq \mu\left(\left\{\mathbf{v} \geq 0 \;\middle|\; \|\mathbf{v}\|_\infty \geq 1, \frac{\mathbf{v}}{\|\mathbf{v}\|_\infty} \in \cdot\right\}\right)$ the angular measure defined on S_∞^{d-1}, the faces

$$\Omega_{\alpha, i_0} \coloneqq \left\{\mathbf{v} \in \Omega_\alpha \;\middle|\; v^{i_0} = 1, \mathbf{v}^{|\alpha \setminus \{i_0\}} < 1\right\}$$

are shown to contain all of Φ's mass, and the restrictions Φ_{α, i_0} of Φ to Ω_{α, i_0} to be absolutely continuous with respect to $dx_{\alpha \setminus \{i_0\}}$, and therefore to have densities $\frac{d\Phi_{\alpha, i_0}}{dx_{\alpha \setminus \{i_0\}}}$. The last assumption can then be written as follows.

Assumption 3. The angular density is uniformly bounded, so that there exists $M > 0$ verifying

$$\sum_{\substack{\beta \subset \{1,\dots,d\} \\ |\beta| \geq 2}} \sup_{i \in \beta} \sup_{\Omega_{\beta, i}} \frac{d\Phi_{\alpha, i_0}}{dx_{\alpha \setminus \{i_0\}}} < M. \tag{3}$$

Since the densities are hardly computable, this hypothesis appears challenging to verify in practice. However, unbounded probability density functions are quite pathological, making assumption 3 very reasonable within our context.

III. Main theoretical results

III.1 Nonparametric estimation of $\mathcal{M}$

Considering the $\mathbf{V}_i$s cannot be directly computed from the (unknown) true marginals, the most standard approach is to estimate the latter through the empirical cumulative distribution functions and compute the corresponding rank transform, giving us estimates denoted as $\widehat{\mathbf{V}}_i$. In a less general setting, when there is a strongly motivated parametric model on the marginals of the data, it may be beneficial to investigate the influence of a parametric estimation of the cdf, although that is outside the scope of the paper. Denoting $\widehat{\mathbb{P}}_n$ the empirical distribution of the $\widehat{\mathbf{V}}_i$ and following (1), it is most natural to estimate μ with

$$\widehat{\mu}_n(\cdot) \coloneqq \frac{n}{k_n}\widehat{\mathbb{P}}_n\left(\frac{n}{k_n}\cdot\right),$$

where $k_n \in \mathbb{N}$ is such that $k_n \underset{n\to\infty}{\longrightarrow} \infty$ and $\frac{n}{k_n} \underset{n\to\infty}{\longrightarrow} \infty$. To proceed with the estimation of μ specifically on the cones $\mathcal{C}_\alpha$, one cannot directly apply the above estimate to such sets as except for $\alpha = \{1,\dots,d\}$, they almost surely contain no sample and are not necessarily continuity sets of the exponent measure as discussed in section II. Thus, one has to consider slighlty larger sets with nonzero volume, that is the previously introduced R_α^ε. This gives the main construction of the paper, being the estimator

$$\widehat{\mathcal{M}}(\alpha) \coloneqq \widehat{\mu}_n(R_\alpha^\varepsilon).$$

The latter is easily computed as

$$\widehat{\mathcal{M}}(\alpha) = \frac{1}{k_n}\sum_{i=1}^{k_n} \mathbb{1}\left\{\widehat{\mathbf{V}}_{\sigma(i)}^{|\alpha} \geq \frac{n}{k_n}\varepsilon, \widehat{\mathbf{V}}_{\sigma(i)}^{|\alpha^c} < \frac{n}{k_n}\varepsilon\right\}, \tag{4}$$

where σ sorts the original data for the infinity norm, i.e. $\left\|\mathbf{X}_{\sigma(1)}\right\|_\infty \geq \cdots \geq \left\|\mathbf{X}_{\sigma(n)}\right\|_\infty$. The authors then go on to prove non-asymptotic error bounds so as to confirm the validity of this approach and get information about the rate of convergence, in a way we now sketch.

III.2 Bounding the error $\left\|\widehat{\mathcal{M}} - \mathcal{M}\right\|_\infty$

The first remark is that the triangle inequality gives an upper bound on the error as the sum of the error related to the estimation of μ with $\widehat{\mu}_n$ and the bias introduced by the use of R_α^ε instead of $\mathcal{C}_\alpha$, namely:

$$\left\|\widehat{\mathcal{M}} - \mathcal{M}\right\|_\infty \leq \max_\alpha \left|\mu - \widehat{\mu}_n\right|(R_\alpha^\varepsilon) + \max_\alpha \left|\mu(\mathcal{C}_\alpha) - \mu(R_\alpha^\varepsilon)\right|, \tag{5}$$

where we denote $\left|\mu - \widehat{\mu}_n\right|(\cdot) \coloneqq \left|\mu(\cdot) - \widehat{\mu}_n(\cdot)\right|$ for brevity. The method employed in the paper is then to bound each term separately. For the first one, the author extends bounds they have developed in [2] to a class of rectangles extending both the R_α^εs and the $[\mathbf{0},\mathbf{v}]^c$s. Intuitively, this class allows the values of ε to differ based on the coordinate, and allows the two considered sets of coordinates to be non complementary, which is formalized as

$$R(\mathbf{x},\mathbf{z},\alpha,\beta) \coloneqq \left\{\mathbf{y} \in [\mathbf{0},\boldsymbol{\infty}]^d, \mathbf{y}^{|\alpha} \geq \mathbf{x}^{|\alpha}, \mathbf{y}^{|\beta} < \mathbf{z}^{|\beta}\right\}. \tag{6}$$

In the same spirit, one defines a "generalized cumulative distribution function" of $\mathbf{U} \coloneqq \mathbf{V}^{-1}$ (which has marginals $\mathcal{U}[0,1]$ under assumption 1) as

$$\tilde{F}_{\alpha,\beta}(\mathbf{x},\mathbf{z}) = \mathbb{P}\left(\mathbf{U} \in R(\mathbf{x},\mathbf{z},\alpha,\beta)\right), \tag{7}$$

and the associated "extreme" version

$$g_{\alpha,\beta}(\mathbf{x},\mathbf{z}) \coloneqq \lim_{t\to\infty} \tilde{F}_{\alpha,\beta}\left(t^{-1}\mathbf{x}, t^{-1}\mathbf{z}\right) \tag{8}$$

where "extreme" is understood over $\mathbf{V}$, whence the factor t^{-1} when considering $\mathbf{U}$. In that way, we have from the regular variation

$$g_{\alpha,\beta}(\mathbf{x},\mathbf{z}) = \mu\left(R\left(\mathbf{x}^{-1},\mathbf{z}^{-1},\alpha,\beta\right)\right). \tag{9}$$

Then, when considering the natural empirical version $\widehat{g}_{n,\alpha,\beta}$ of $g_{\alpha,\beta}$ from (8) (that is, by computing the probability defining $\tilde{F}_{\alpha,\beta}$ with the empirical distribution and using an order statistic instead of t), one actually recovers that

$$\widehat{g}_{n,\alpha,\beta} = \widehat{\mu}_n\left(R\left(\mathbf{x}^{-1},\mathbf{z}^{-1},\alpha,\beta\right)\right). \tag{10}$$

Therefore, it becomes clear that uniformly bounding the error between the $\widehat{g}_{n,\alpha,\beta}$s and the $g_{\alpha,\beta}$s will yield a bound on the first term of (5), which allows the authors to extend their previous work in [2] to obtain new bounds. Writing $\tilde{\boldsymbol{\varepsilon}}$ such that $\tilde{\boldsymbol{\varepsilon}}^{|\alpha} = \mathbf{1}^{|\alpha}$, $\tilde{\boldsymbol{\varepsilon}}^{|\alpha^c} = \boldsymbol{\varepsilon}^{|\alpha^c}$, one can recover the desired rectangles as

$$R_\alpha^\varepsilon = R(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}, \alpha, \alpha^c) \setminus R(\boldsymbol{\varepsilon}, \tilde{\boldsymbol{\varepsilon}}, \alpha, \{1, \dots, d\}), \tag{11}$$

and from the triangle inequality and the fact that for $\varepsilon < 1$, $\tilde{\boldsymbol{\varepsilon}} \geq \boldsymbol{\varepsilon}$, we have

$$\begin{aligned} |\mu - \widehat{\mu}_n| \left(R_\alpha^\varepsilon\right) &\leq |\mu - \widehat{\mu}_n| \left(R(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}, \alpha, \alpha^c)\right) + |\mu - \widehat{\mu}_n| \left(R(\boldsymbol{\varepsilon}, \tilde{\boldsymbol{\varepsilon}}, \alpha, \{1, \dots, d\})\right) \\ &\leq 2 \max_\beta \sup_{\varepsilon \leq \mathbf{x}, \mathbf{z}} |\mu - \widehat{\mu}_n| \left(R(\mathbf{x}, \mathbf{z}, \alpha, \beta)\right). \end{aligned} \tag{12}$$

From (9) and (10), this last term can be rewritten in terms of the $g_{\alpha,\beta}$s and their empirical counterparts, and using a slightly altered version of the VC-type bounds in [2], one can bound the first term of (5): there exists $C > 0$ such that for $0 < \varepsilon < \frac{1}{4}$, $\delta \geq e^{-k_n}$, with probability at least $1 - \delta$,

$$\max_\alpha |\mu - \widehat{\mu}_n| \left(R_\alpha^\varepsilon\right) \leq Cd \sqrt{\frac{1}{\varepsilon k_n} \ln\left(\frac{d+3}{\delta}\right)} + 2 \max_{\alpha,\beta} \sup_{\mathbf{0} \leq \mathbf{x}, \mathbf{z} \leq 2\varepsilon^{-1}} \left| \frac{n}{k_n} \tilde{F}_{\alpha,\beta}\left(\frac{k_n}{n}\mathbf{x}, \frac{k_n}{n}\mathbf{z}\right) - g_{\alpha,\beta}(\mathbf{x}, \mathbf{z}) \right|. \tag{13}$$

Note that the assumptions on δ is quickly satisfied when k_n grows, meaning one can easily obtain a bound holding with probability almost 1.

To bound the second term in (5), the authors use their assumptions 2 and 3 to show through technicalities that

$$|\mu(R_\alpha^\varepsilon) - \mu(\mathcal{C}_\alpha)| \leq Md^2\varepsilon. \tag{14}$$

This allows one to have convergence towards 0 proportional to ε, although said proportional factor can quickly be large when the dimension increases.

Thanks to the previous bounds (5), (13) and (14), the main result of the paper can be stated:

Theorem 1. *Under assumptions 2 and 3, there exists $C < 0$, such that for $0 < \varepsilon < \frac{1}{4}$, $\delta \geq e^{-k_n}$, with probability at least $1 - \delta$,*

$$\left\| \widehat{\mathcal{M}} - \mathcal{M} \right\|_\infty \leq Cd \left(\sqrt{\frac{1}{\varepsilon k_n} \ln\left(\frac{d+3}{\delta}\right)} + Md\varepsilon \right) + 2 \max_{\alpha,\beta} \sup_{\mathbf{0} \leq \mathbf{x}, \mathbf{z} \leq 2\varepsilon^{-1}} \left| \frac{n}{k_n} \tilde{F}_{\alpha,\beta}\left(\frac{k_n}{n}\mathbf{x}, \frac{k_n}{n}\mathbf{z}\right) - g_{\alpha,\beta}(\mathbf{x}, \mathbf{z}) \right|. \tag{15}$$

Note that the second term is a slightly different bound to that stated in the paper, but it follows more directly from the above, and how the authors removed the maximum over β in favor of α^c was not made explicit. Although, it ultimately makes little difference since the term written in (15) is just as easily shown to go to zero following [3] and taking the maximum over a finite class. It is a looser bound, but since neither have explicit convergence rates, they basically give the same information. We now discuss this result and the underlying methodology.

The first term of the bound (15) quantifies the trade-off in ε: reducing its value by an order of magnitude shrinks the "ε-thickening bias" the same amount but increases the estimate error by at least half an order of magnitude. Under this light, the "standard" value of $\varepsilon = 0.01$ suggested by the authors in the numerical section discussed in section V can be seen as a choice to multiply the estimate error by not much more than one order of magnitude. As highlighted by the authors, the dependency in $\mathcal{O}\left(\frac{1}{\sqrt{k_n}}\right)$ is not too surprising given classical VC inequalities and the fact that only k_n (extreme) samples are effectively counted to estimate the mass on truncated cones. The second term however, is much less informative in regards to convergence speed. But given its derivation, one can imagine this is in practice a rather loose bound, leaving hope of reasonably fast convergence.

The assumptions 2 and 3 are only used by the authors to bound the ε-thickening error, as the framework developed in [2] needs no such assumptions. While the latter end up showing guidance for parameter choice as dicussed above and further explored in section V, one can see that convergence will still hold should these assumptions fail to be verified.

III.3 Thresholding $\widehat{\mathcal{M}}$

Considering data is in practice noisy, one will often obtain some αs for which $\widehat{\mathcal{M}}(\alpha)$ is nonzero but negligible, so that one could assume the corresponding $\mathcal{M}(\alpha)$ is null. To choose only the meaningful directions, the authors propose removing values of $\widehat{\mathcal{M}}(\alpha)$ under a threshold computed as $p \left| \left\{ \alpha \;\middle|\; \widehat{\mathcal{M}}(\alpha) > 0 \right\} \right|^{-1} \sum_\alpha \widehat{\mathcal{M}}(\alpha)$ i.e. some proportion $p > 0$ of the average mass of

faces with positive mass. Denoting $\tilde{\mathcal{M}}$ the estimator obtained via this thresholding operation, its deviation from $\widehat{\mathcal{M}}$ is by definition at most the threshold, which means that

$$\left\|\tilde{\mathcal{M}} - \mathcal{M}\right\|_\infty \leq \left\|\widehat{\mathcal{M}} - \mathcal{M}\right\|_\infty + p\left|\left\{\alpha \;\middle|\; \widehat{\mathcal{M}}(\alpha) > 0\right\}\right|^{-1} \sum_\alpha \widehat{\mathcal{M}}(\alpha). \tag{16}$$

Such a thresholding therefore adds an error term proportional to p, so that we recover convergence when $p \to 0$. In fact, the authors show that $\widehat{\mathcal{M}}(\alpha)$ can be seen as an empirical risk minimizer, corresponding to the natural L^2 loss considering what we are estimating, namely

$$\mathcal{R} : \left| \begin{array}{ccc} \mathbb{R}^{2^d-1} & \longrightarrow & \mathbb{R}_+ \\ \mathbf{m} & \longmapsto & \mathbb{E}\left[\left\|\mathbb{1}\left\{\mathbf{V} \in \frac{k_n}{n} R_\alpha^\varepsilon\right\}\mathbf{1} - \mathbf{m}\right\|_2 \;\middle|\; \|V\|_\infty \geq \frac{n}{k_n}\right] \end{array} \right. .$$

Thus, the thresholding can be seen as L^1 regularization of the above risk to obtain a more sparse representation.

IV. Anomaly detection

With the estimator $\widehat{\mathcal{M}}$, the authors introduce a scoring function Detecting Anomalies among Multivariate EXtremes (DAMEX). We denote in the following T the rank transform giving $\mathbf{V}$ from $\mathbf{X}$, $\widehat{T}$ the empirical version as discussed in III.1, and $\alpha(\mathbf{x})$ the unique subset of $\{1, \dots, d\}$ such that $\mathbf{x} \in R_\alpha^\varepsilon$. The underlying idea is more or less statistical hypothesis testing: given a new extreme point $\mathbf{x}$ (where extreme is quantified by $T(\mathbf{x}) \geq \frac{k_n}{n}$), one can estimate (up to a constant) the probability under the "normal" distribution learned by the estimator of lying in a more extreme region in the same direction as $\mathbf{x}$, in the spirit of a p-value. Formally, define the *directional tail region* corresponding to $\mathbf{x}$ as

$$A_\mathbf{x} \coloneqq \left\{\mathbf{y} \;\middle|\; T(\mathbf{y}) \in R_{\alpha(\mathbf{x})}^\varepsilon, \|T(\mathbf{y})\|_\infty \geq \|T(\mathbf{x})\|\right\}.$$

Then, we have

$$\begin{aligned} \mathbb{P}(\mathbf{X} \in A_\mathbf{x}) &= \mathbb{P}(\mathbf{V} \in \|T(\mathbf{x})\|_\infty R_{\alpha(\mathbf{x})}^\varepsilon) \\ &= \mathbb{P}(\|\mathbf{V}\|_\infty \geq \|T(\mathbf{x})\|_\infty)\mathbb{P}(\mathbf{V} \in \|T(\mathbf{x})\|_\infty R_{\alpha(\mathbf{x})}^\varepsilon \mid \|\mathbf{V}\|_\infty \geq \|T(\mathbf{x})\|_\infty) \\ &= \underbrace{\mathbb{P}(\|\mathbf{U}\|_\infty \leq \|T(\mathbf{x})\|_\infty^{-1})}_{=\|T(\mathbf{x})\|_\infty^{-1} \text{ (assumption 1)}} \underbrace{\mathbb{P}(\mathbf{V} \in \|T(\mathbf{x})\|_\infty R_{\alpha(\mathbf{x})}^\varepsilon \mid \|\mathbf{V}\|_\infty \geq \|T(\mathbf{x})\|_\infty)}_{\substack{\longrightarrow \frac{\mathcal{M}(\alpha(\mathbf{x}))}{\mu([\mathbf{0},\mathbf{1}]^c)} \\ \|T(\mathbf{x})\|_\infty \to \infty \\ \varepsilon \to 0}}, \end{aligned}$$

which motivates the scoring function

$$\widehat{s}(\mathbf{x}) \coloneqq \frac{\widehat{\mathcal{M}}(\alpha(\mathbf{x}))}{\left\|\widehat{T}(\mathbf{x})\right\|_\infty}. \tag{17}$$

Note that this score ignores the constant $\mu([\mathbf{0}, \mathbf{1}]^c)$ since it is unknown, meaning that the interpretation as a probability is lost, and thus choosing a threshold under which to consider a data point an anomaly is not interpretable as the choice of a p-value threshold despite the analogy. The choice of such a threshold being application-dependent, the preferred metrics to evaluate the performance of the algorithm are the ROC AUC and PR AUC. We now use these in a comparative study on real data.

V. Numerical experiments

The DAMEX pipeline was reimplemented in the supplementary notebook [4], where comparisons are made with iForest [5] as what is made in [1], but also with Local Outlier Factor (LOF) [6]. The influence of k_n and ε are also investigated in more detail.

We consider the same `forestcover` dataset as in the paper, obtained from [7]. This dataset has 54 features and contains labels separating normal and anormal data. One should note that these features are encoded as integers and contain some duplicate values, meaning the assumptions used to derive bounds do not hold as discussed in section II. Nevertheless, convergence will still hold as seen in part III.2. We follow the authors' methodology, being that the training set is only composed of normal entries, and testing is only made in the extreme region (i.e. $\left\{\mathbf{x} \;\middle|\; \left\|\widehat{T}(\mathbf{x})\right\|_\infty \geq k_n\right\}$). Over 10 experiments and for a subset

of $N = 80000$ samples, we take 80% of the normal data as training set ($n = \lfloor 0.8N \rfloor$), and we look at the average ROC AUC & PR AUC over a test set consisting of the extremes present in the remaining 20% of normal data and the anormal data. We additionally observe the average dimensions of the faces of nonzero estimated mass, i.e. the (uniform) average of $\left\{|\alpha|, \widehat{\mathcal{M}}(\alpha) > 0\right\}$, shortened as AFD (Average Face Dimension). We make the expression of k_n vary with fixed $\varepsilon = 0.01$, $p = 0.1$ and obtain the results shown in table 1.

	DAMEX			IsolationForest		LocalOutlierFactor	
k_n	AUC ROC	AUC PR	AFD	AUC ROC	AUC PR	AUC ROC	AUC PR
$n^{\frac{1}{4}}$	0.503	0.054	8.76	0.947	0.496	**0.996**	**0.961**
$\sqrt{n}$	0.895	0.678	24.6	0.884	0.614	**0.994**	**0.982**
$n^{\frac{3}{4}}$	0.817	0.773	54.0	0.715	0.498	**0.994**	**0.987**
$n^{\frac{1}{4}}\ln(n)$	0.939	0.806	26.2	0.911	0.638	**0.994**	**0.981**

Table 1: Results on extreme region with varying k_n, $\varepsilon = 0.01$

Despite the rather high dimensionality of the problem, LOF very convincingly surpasses both algorithms compared in the paper. The results show the trade-off in k_n: low values lead to underfitting, while too high values mean the considered order statistics contain non-extreme data and thus fail to recover a sparse representation (high AFD).A good compromise seems to be reached when taking $k_n = n^{\frac{1}{4}}\ln(n)$ for this dataset, although it is hardly interpretable through (15). Some sparsity can be observed as highlighted in [1] on other datasets, as for reasonable values of k_n, the AFD is less than half of the initial dimensionality. We also note that despite a suboptimal implementation using python, the run times were acceptable and not awfully longer than LOF. This highlights the relatively moderate ($\mathcal{O}(dn\ln(n))$) complexity of DAMEX. To get a more adaptative heuristic for the choice of ε, consider the first term of (15). Following the remark made by the authors that if the angular density is constant one has $M \leq d$, and forgetting the logarithmic term for simplicity, let us try to choose ε minimizing the quantity $\frac{1}{\sqrt{\varepsilon k_n}} + d^2\varepsilon$. This yields $\varepsilon = \frac{\sqrt{k_n}}{d^{\frac{4}{3}}}$, which unfortunately goes to $+\infty$ when $n \to \infty$. Nevertheless, we now evaluate on the same dataset and with our best value of $k_n = n^{\frac{1}{4}}\ln(n)$ but increasing the number N of total data points considered to see if there is some relevance to this approach.

	DAMEX			IsolationForest		LocalOutlierFactor	
N	AUC ROC	AUC PR	AFD	AUC ROC	AUC PR	AUC ROC	AUC PR
80000	0.924	0.711	20.9	0.873	0.551	0.994	0.981
150000	0.906	0.639	20.7	0.890	0.600	0.994	0.981

Table 2: Results on extreme region with varying N, $k_n = n^{\frac{1}{4}}\ln(n)$, $\varepsilon = \frac{(k_n)^{\frac{1}{3}}}{d^{\frac{4}{3}}}$

It would appear such a choice of ε yields decent results on this dataset, although performance decreases with N as expected. This could potentially be an alternative starting value if $\varepsilon = 0.01$ gives poor performance, in the case where N is not too large in a high-dimensional setting, but further investigation is beyond the scope of this report.

VI. Conclusion

The paper studied in this report has for main contribution the statistical estimator of the angular measure over relevant subcones along with convergence bounds, capturing potential sparsity patterns among extreme dependencies so as to reduce dimensionality and bridge the gap with low-dimensional methods. The second term in said convergence bound (15) could potentially be made more informational if one were to make additional regularity assumptions, e.g. that it is itself regularly varying. That could allow for more informed parameter choosing, considering that a cross validation as in section V can be computationally costly and the explored heuristic choice of ε is unsatisfying from a theoretical standpoint. Although, said assumptions should be made carefully considering even the initial hypotheses of this paper do not hold for a nonnegligible class of datasets. Another potential way to expand on the paper's method would consist in utilizing this dimensional reduction procedure as a preprocessing step before applying standard algorithms to the exhibited subcones. That is, in the suggested application to anomaly detection, one could increase the accuracy of the method by not only flagging extreme among cones of negligible angular mass as anomalies, but additionally allow detection of anomalies within normal, lower dimensional cones through other methods. Finally, one can see that the choice of representation has high impact on the method: the considered subcones correspond to sets of nonzero coordinates. Thus, the combination of the paper's method with different choices of basis and e.g. manifold representations of data could be investigated.

References

[1] Nicolas Goix, Anne Sabourin, and Stephan Clémençon. "Sparse representation of multivariate extremes with applications to anomaly detection". In: *Journal of Multivariate Analysis* 161 (2017), pp. 12–31. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2017.06.010. URL: https://www.sciencedirect.com/science/article/pii/S0047259X17304062.

[2] N. Goix, A. Sabourin, and S. Clémençon. "Learning the dependence structure of rare events: a non-asymptotic study". In: *Proc. COLT*. 2015.

[3] Yongcheng Qi. "Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics". English (US). In: *Acta Mathematicae Applicatae Sinica* 13.2 (1997), pp. 167–175. ISSN: 0168-9673. DOI: 10.1007/BF02015138.

[4] M. Hardion. *damex notebook*. URL: https://github.com/mhardion/damex.

[5] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation Forest". In: *2008 Eighth IEEE International Conference on Data Mining*. 2008, pp. 413–422. DOI: 10.1109/ICDM.2008.17.

[6] Markus M. Breunig et al. "LOF: Identifying Density-Based Local Outliers". In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. SIGMOD '00. Dallas, Texas, USA: Association for Computing Machinery, 2000, pp. 93–104. ISBN: 1581132174. DOI: 10.1145/342009.335388. URL: https://doi.org/10.1145/342009.335388.

[7] The scikit-learn community. *Forest covertypes dataset*. scikit-learn 1.3.2 documentation. URL: https://scikit-learn.org/stable/datasets/real_world.html#covtype-dataset (visited on 03/12/2023).