A Multivariate Extreme Value Theory Approach to Anomaly Clustering and Visualization

Maël Chiapino, Stéphan Clémençon, Vincent Feuillard, Anne Sabourin

A Multivariate Extreme Value Theory Approach to Anomaly Clustering and Visualization

Maël Chiapino[1], Stephan Clémençon[1], Vincent Feuillard[2], and Anne Sabourin[1]

[1] LTCI, Télécom Paris, Institut polytechnique de Paris, France,
anne.sabourin@telecom-paristech.fr
[2] Airbus Central R&T, AI Research

July 16, 2019

Abstract

In a wide variety of situations, anomalies in the behaviour of a complex system, whose health is monitored through the observation of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ valued in $\mathbb{R}^d$, correspond to the simultaneous occurrence of extreme values for certain subgroups $\alpha \subset \{1, \ldots, d\}$ of variables X_j. Under the heavy-tail assumption, which is precisely appropriate for modeling these phenomena, statistical methods relying on multivariate extreme value theory have been developed in the past few years for identifying such events/subgroups. This paper exploits this approach much further by means of a novel mixture model that permits to describe the distribution of extremal observations and where the anomaly type α is viewed as a latent variable. One may then take advantage of the model by assigning to any extreme point a posterior probability for each anomaly type α, defining implicitly a similarity measure between anomalies. It is explained at length how the latter permits to cluster extreme observations and obtain an informative planar representation of anomalies using standard graph-mining tools. The relevance and usefulness of the clustering and 2-d visual display thus designed is illustrated on simulated datasets and on real observations as well, in the aeronautics application domain.

Keywords— Anomaly detection, clustering, graph-mining, latent variable analysis, mixture modelling, multivariate extreme value theory, visualization

1 Introduction

Motivated by a wide variety of applications ranging from fraud detection to aviation safety management through the health monitoring of complex net-

works, data center infrastructure management or food risk analysis, unsupervised anomaly detection is now the subject of much attention in the data science literature, see *e.g.* Gorinevsky et al. (2012); T. Fawcett (1997); Viswanathan et al. (2012). In frequently encountered practical situations and from the viewpoint embraced in this paper, anomalies coincide with rare measurements that are extremes, *i.e.* located far from central statistics such as the sample mean. In the 1-d setting, numerous statistical techniques for anomaly detection are based on a parametric representation of the tail of the observed univariate probability distribution, relying on *extreme value theory* (EVT), see *e.g.* Clifton et al. (2011); Lee and Roberts (2008); Roberts (2000); Tressou (2008) among others. In (even moderately) large dimensional situations, the modelling task becomes much harder. Many nonparametric heuristics for supervised classification have been adapted, substituting rarity for labeling, see *e.g.* Schölkopf et al. (2001), Steinwart et al. (2005) or Liu et al. (2008). In the unsupervised setting, several extensions of the basic linear Principal Component Analysis for dimensionality reduction and visualization techniques have been proposed in the statistics and data-mining literature, accounting for non linearities or increasing robustness for instance, *cf* Gorban et al. (2008) and Kriegel et al. (2008). These approaches intend to describe parsimoniously the 'center' of a massive data distribution, see *e.g.* Naik (2017) and the references therein. Similarly, for clustering purposes, several multivariate heavy-tailed distributions have been proposed that are robust to the presence of outliers, see *e.g.* Forbes and Wraith (2014), Punzo and Tortora (2018). However the issue of clustering extremes or outliers is only recently receiving attention, at the instigation of industrial applications such as those mentioned above and because of the increasing availability of extreme observations in databases: generally out-of-sample in the past, extreme values are becoming observable in the Big Data era. It is the goal of the present article to propose a novel mixture model-based approach for clustering extremes in the multivariate setup, *i.e.* when the observed random vector $\mathbf{X} = (X_1, \ldots, X_d)$ takes its values in the positive orthant of the space $\mathbb{R}^d$ with $d > 1$ equipped with the sum-norm $\|(x_1, \ldots, x_d)\| = \sum_{1 \leq j \leq d} |x_j|$: 'extremes' coinciding then with values x such that $\mathbb{P}(\|\mathbf{X}\| > \|\mathbf{x}\|)$ is 'extremely small'. Precisely, it relies on a dimensionality reduction technique of the tail distribution recently introduced in Goix et al. (2017) and Goix et al. (2016), and referred to as the DAMEX algorithm. Based on multivariate extreme value theory (MEV theory), the latter method may provide a hopefully sparse representation of the support of the angular measure related to the supposedly heavy-tailed distribution of the random vector $\mathbf{X}$. As the angular measure asymptotically describes the dependence structure of the variables X_j in the extremal domain (and, roughly speaking, permits to assign limit probabilities to directions $\mathbf{x}/\|\mathbf{x}\|$ in the unit sphere along which extreme observations may occur), this statistical procedure identifies the groups $\alpha \subset \{1, \ldots, d\}$ of feature indices such that the collection of variables $\{X_j : j \in \alpha\}$ may be simultaneously very large, while the others, the X_j's for $j \notin \alpha$, remain small. Groups of this type are in 1-to-1 correspondence with the faces $\Omega_\alpha = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1, x_j = 0 \text{ if } j \notin \alpha \text{ and } x_j > 0 \text{ if } j \in \alpha\}$ of the unit sphere composing the support of the angular measure. In practice, a sparse

representation of the extremal dependence structure is obtained when only a few such groups of variables can be exhibited (compared to $2^d - 1$) and/or when these groups involve a small number of variables (with respect to d). Here we develop this framework further, in order to propose a (soft) clustering technique in the region of extremes and derive effective 2-d visual displays, sheding light on the structure of anomalies/extremes in sparse situations. This is achieved by modelling the distribution of extremes as a specific *mixture model*, where each component generates a different type α of extremes. In this respect, the present paper may be seen as an extension of Boldi and Davison (2007); Sabourin and Naveau (2014), where a Bayesian inference framework is designed for moderate dimensions ($d \leq 10$ say) and situations where the sole group of variables with the potential of being simultaneously large is $\{1, \ldots, d\}$ itself. In the context of mixture modelling (see *e.g.* Fruhwirth-Schnatter et al. (2018)), the Expectation-Maximization algorithm (EM in abbreviated form) permits to partition/cluster the set of extremal data through the statistical recovery of *latent observations*, as well as posterior probability distributions (inducing a soft clustering of the data in a straighforward manner) and, as a by-product, a similarity measure on the set of extremes: the higher the probability that their latent variables are equal, the more similar two extreme observations X and X' are considered. The similarity matrix thus obtained naturally defines a *weighted graph*, whose vertices are the anomalies/extremes observed, paving the way for the use of powerful graph-mining techniques for community detection and visualization, see *e.g.* Schaeffer (2007), Hu and Shi (2015) and the references therein. Beyond its detailed description, the methodology proposed is applied to a real fleet monitoring dataset in the aeronautics domain and shown to provide useful tools for analyzing and interpreting abnormal data.

The paper is structured as follows. Basic concepts of MEV theory are briefly recalled in Section 2, in particular the concept of angular measure, together with the technique proposed in Goix et al. (2016, 2017) for estimating the (hopefully sparse) support of the latter, which determines the dependence structure of extremes arising from a heavy-tailed distribution. Section 3 details the mixture model we propose to describe the distribution of extreme data, based on the output of the support estimation procedure, together with the EM algorithm variant we introduce in order to estimate its parameters. It is next explained in Section 5 how to exploit the results of this inference method to define a similarity matrix of the extremal data, reflecting a weighted graph structure of the observed anomalies, and apply dedicated community detection and visualization techniques so as to extract meaningful information from the set of extreme observations. The relevance of the approach we promote is finally illustrated by numerical experiments, on synthetic and real data in Section 6. An implementation of the proposed method and the code for the experiments carried out in this paper are available online[1]. Technical details are deferred to the Appendix section.

[1] https://github.com/mchiapino/mevt_anomaly

2 Background and Preliminaries

We start with recalling key notions of MEVT, the concept of angular measure in particular, as well as the inference method investigated in Goix et al. (2016, 2017) to estimate its support. Here and throughout, the Dirac mass at any point x is denoted by δ_x, the indicator function of any event A by $\mathbb{1}\{A\}$, the cardinality of any finite set E by $|E|$. Capital letters generally refer to random quantities whereas lower case ones denote deterministic values. Finally, boldface letters denote vectors as opposed to Roman letters denoting real numbers.

2.1 Heavy-Tail Phenomena - Multivariate Regular Variation

Extreme Value Theory (EVT) describes phenomena that are not governed by an 'averaging effect' but can be instead significantly impacted by very large values. By focusing on large quantiles rather than central statistics such as the median or the sample mean, EVT provides models for the unusual rather than the usual and permits to assess the probability of occurence of rare (extreme) events. Application domains are numerous and diverse, including any field related to risk management as finance, insurance, environmental sciences or aeronautics. Risk monitoring is a typical use case of EVT. The reader is referred to Coles (2001) and the references therein for an introduction to EVT and its applications. In the univariate setting, typical quantities of interest are high quantiles of a random variable X, *i.e.* $1-p$ quantiles for $p \to 0$. When p is of the same order of magnitude as $1/N$ or smaller, empirical estimates become meaningless. Another issue is the estimation of the probability of an excess over a high threshold u, $p_u = \mathbb{P}(X > u)$ when few (or none) observations are available above u. In such contexts, EVT essentially consists in using a parametric model (the generalized Pareto distributions) for the tail distribution, which is theoretically justified asymptotically, *i.e.* when $p \to 0$ or $u \to \infty$. Here and throughout we place ourselves in the context where the variable of interest is regularly varying; see Resnick (1987, 2007) for a general introduction to regular variation and its applications to data analysis. In the univariate case the required assumption is the existence of a sequence $a_n > 0$ such that $a_n \to \infty$ and a function $h(x)$ such that $n\mathbb{P}\left(X/a_n > x\right) \xrightarrow[n\to\infty]{} h(x)$, $x > 0$. Notice that this assumption is satisfied by most textbook heavy tailed distributions, *e.g.* Cauchy, Student. In such a case h is necessarily of the form $h(x) = Cx^{-\alpha}$ for some $C, \alpha > 0$, where α is called the *tail index* of X and a_n may be chosen as $a_n = n^{1/\alpha}$. In the multivariate setting, consider a d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$, the goal is to infer quantities such as $\mathbb{P}\left(X_1 > x_1, \ldots, X_d > x_d\right)$ for large $x_1, \ldots, x_d$. A natural first step is to standardize each marginal distribution so that the X_j's are all regularly varying with tail index $\alpha = 1$ and scaling constant $C = 1$. One convenient choice is to use the probability integral transform. For $\mathbf{x} = (x_1, \ldots, x_d)$, let $F_j(x_j) = \mathbb{P}\left(X_j \leq x_j\right)$. Assuming that F_j is continuous, the transformed variable $V_j = (1 - F_j(X_j))^{-1}$ follows a Pareto distribution, $\mathbb{P}\left(V_j > v\right) = v^{-1}$, $v \geq 1$.

In practice F_j is unknown but its empirical version $\hat{F}_j$ may be used instead. Another option when each X_j is regularly varying with tail index α_j is to estimate (α_j, C_j) using *e.g.* a generalized Pareto model above large thresholds, see Coles (2001) or Beirlant et al. (2004) and the references therein. Then $V_j = X^{\alpha_j}/C_j$ is standard regularly varying, meaning that $n\mathbb{P}\left(V_j/n > x\right) \to x^{-1}, x > 0$ as $n \to \infty$. The multivariate extension of the latter assumption is that the standardized vector $\mathbf{V} = (V_1, \ldots, V_d)$ is regularly varying with tail index equal to 1, *i.e.* there exists a limit Radon measure μ on $\mathbb{R}_+^d \setminus \{0\}$ such that

$$n\mathbb{P}\left(n^{-1}\mathbf{V} \in A\right) \to \mu(A) \tag{1}$$

for all A in the continuity set of μ such that $0 \notin \partial A$. The measure μ is called the *exponent measure*. In the standard setting characterized by (1), it is homogeneous of order -1, that is $\mu(tA) = t^{-1}\mu(A)$, where $tA = \{t\mathbf{v}, \mathbf{v} \in A\}, A \subset \mathbb{R}_d^+$ and $\mu\{\mathbf{x} \in \mathbb{R}_+^d : x_j \geq 1\} = 1$. Assumption (1) applies immediately to the problem of estimating the probability of reaching a set tA which is far form $\mathbf{0}$ (*i.e.* t is large): one may write $\mathbb{P}(\mathbf{V} \in tA) \approx \frac{1}{t}\mu(A)$, so that estimates of μ automatically provide estimates for such quantities. In a word, μ may be used to characterize the distributional tail of $\mathbf{V}$. For modeling purposes, the homogeneity property $\mu(t\cdot) = t^{-1}\mu(\cdot)$ suggests a preliminary decomposition of μ within a (pseudo)-polar coordinates system, as detailed next.

2.2 Angular Measure - Dependence in the Extremes

Consider the sum-norm $\|\mathbf{v}\| := v_1 + \ldots + v_d$ and $\mathcal{S}_d := \{\mathbf{w} \in \mathbb{R}_+^d : \|\mathbf{w}\| = 1\}$ the d-dimensional simplex. Introduce the polar transformation $T : \mathbf{v} \mapsto T(\mathbf{v}) = (r, \mathbf{w})$ defined on $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$, where $r = \|\mathbf{v}\|$ is the radial component and $\mathbf{w} = r^{-1}\mathbf{v}$ is the angular one. Now define the *angular measure* Φ on $\mathcal{S}_d$ (see *e.g.* Resnick (2007) or Beirlant et al. (2004) and the references therein): $\Phi(A) := \mu\left\{\mathbf{v} : \|\mathbf{v}\| > 1, \|\mathbf{v}\|^{-1}\mathbf{v} \in A\right\}\}$, with $A \subset \mathcal{S}_d$. Notice that $\Phi(\mathcal{S}_d) < \infty$ and, by homogeneity,

$$\mu \circ T^{-1}(\mathrm{d}r, \mathrm{d}\mathbf{w}) = r^{-2}\mathrm{d}r\Phi(\mathrm{d}\mathbf{w}). \tag{2}$$

In other words the exponent measure μ factorizes into a tensor product of a radial component and an angular component. Setting $R = \|\mathbf{V}\|$ and $\mathbf{W} = R^{-1}\mathbf{V}$, a consequence is that

$$\mathbb{P}\left(\mathbf{W} \in A, R > tr \mid R > t\right) \xrightarrow[t\to\infty]{} r^{-1}\Phi(\mathcal{S}_d)^{-1}\Phi(A) \tag{3}$$

for all measurable set $A \subset \mathcal{S}_d$ such that $\Phi(\partial A) = 0$ and $r > 1$. Hence, given that the radius R is large, R and the angle $\mathbf{W}$ are approximately independent, the distribution of $\mathbf{W}$ is approximately the angular measure – up to a normalizing constant $\Phi(\mathcal{S}_d)$ – and R follows approximately a Pareto distribution. As it describes the distribution of the directions formed by the largest observations,

the angular measure exhaustively accounts for the dependence structure in the extremes. Our choice of a standard regular variation framework (1) and that of the sum-norm yield the following moment constraint on Φ:

$$\int_{\mathcal{S}_d} w_i \, \Phi(\mathrm{d}\mathbf{w}) = 1, \text{ for } i = 1, \ldots, d. \tag{4}$$

In addition, the normalizing constant is explicit:

$$\Phi(\mathcal{S}_d) = \int_{\mathcal{S}_d} \Phi(\mathrm{d}\mathbf{w}) = \int_{\mathcal{S}_d} (w_1 + \ldots + w_d)\Phi(\mathrm{d}\mathbf{w}) = d. \tag{5}$$

Remark 1. *The choice of the sum-norm here is somewhat arbitrary. Any other norm on $\mathbb{R}^d$ for the pseudo-polar transformation is equally possible, leading to alternative moment constraints and normalizing constants. The advantage of the sum-norm is that it allows convenient probabilistic modeling of the angular component $\mathbf{w}$ on the unit simplex.*

2.3 Support Estimation - the DAMEX Algorithm

We now expose the connection between Φ's (or equivalently, μ's) support and the subsets of components which may simultaneously take very large values, while the others remain small.

Sparse support. Fix $\alpha \subset \{1, \ldots, d\}$ and consider the associated truncated cone

$$\mathcal{C}_\alpha = \big\{\mathbf{v} \geq 0 : \ \|\mathbf{v}\|_\infty \geq 1, \ v_i > 0 \text{ for } i \in \alpha, \ v_i = 0 \ \text{ for } i \notin \alpha\big\}. \tag{6}$$

The family $\{\mathcal{C}_\alpha, \alpha \subset \{1, \ldots, d\}, \alpha \neq \emptyset\}$ first introduced in Goix et al. (2016) defines a partition of $\mathbb{R}_+^d \setminus [0,1]^d$ which is of particular interest for our purpose: notice first that, by homogeneity of μ, the following equivalence holds: $\Phi(\mathcal{S}_\alpha) > 0 \Leftrightarrow \mu(\mathcal{C}_\alpha) > 0$, where $\mathcal{S}_\alpha = \big\{\mathbf{v} \in \mathbb{R}_+^d : \ \|v\| = 1, \ v_i > 0 \text{ for } i \in \alpha, v_i = 0 \text{ for } i \notin \alpha\big\}$, $\varnothing \neq \alpha \subset \{1, \ldots, d\}$. Observe next that $\mu(\mathcal{C}_\alpha) > 0$ means that the limiting rescaled probability that 'all features in α are simultaneously large, while the others are small' is non zero. Precisely, consider the ϵ-thickened rectangle

$$\mathcal{R}_\alpha^\epsilon \ = \ \big\{\mathbf{v} \geq 0, \ \|\mathbf{v}\|_\infty \geq 1, \ v_i > \epsilon \ \text{ for } i \in \alpha, \ v_i \leq \epsilon \ \text{ for } i \notin \alpha\big\},$$

which corresponds to the event that all features in α are large, while the other are small. The $\mathcal{R}_\alpha^\epsilon$'s define again a partition of $\mathbb{R}_+^d \setminus [0,1]^d$ for each fixed $\epsilon \geq 0$. In addition, we have that $\mathcal{C}_\alpha = \cap_{\epsilon > 0, \epsilon \in \mathbb{Q}} \mathcal{R}_\alpha^\epsilon$, so that by upper continuity of μ,

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

with

$$\mu(\mathcal{R}_\alpha^\epsilon) = \lim_{t \to \infty} t\mathbb{P}(\|\mathbf{V}\|_\infty > t, \ \forall j \in \alpha : V_j > t\epsilon, \ \ \forall j \notin \alpha : V_j < t\epsilon).$$

In the sequel, set $\mu_\alpha = \mu(\mathcal{C}_\alpha)$, $\mathbb{M} = \big\{\alpha \subset \{1,\dots,d\}, \alpha \neq \varnothing, \mu_\alpha > 0\big\}$. Although every μ_α may be positive in theory, a reasonable assumption in many practical high dimensional situations is that $\mu_\alpha = 0$ for the vast majority of the $2^d - 1$ cones $\mathcal{C}_\alpha$. In other words, not all combinations of coordinates of $\mathbf{V}$ can be large together, so that the support of μ (and that of Φ) is sparse.

Support estimation. The task of estimating the support of μ (or Φ) has recently received increasing attention in the statistics and machine learning literature. Chautru (2015) first proposed a non parametric clustering approach involving principal nested spheres, which provides great flexibility at the price of computational cost. In contrast, Goix et al. (2016)'s methods rely on the above mentioned partition of the unit sphere into $2^d - 1$ sub-simplices Ω_α and led to the so-called DAMEX algorithm which computational complexity $O(dn \log n)$ scales well with higher dimensions. Their algorithm produces the list of α's such that the empirical counterpart of μ_α (denoted $\hat{\mu}_\alpha$ in the sequel) is non zero. Defining a threshold $m_{\min} \geq 0$ below which $\hat{\mu}_\alpha$ is deemed as negligible, one thus obtains a list of subsets $\widehat{\mathbb{M}} = \{\alpha \subset \{1,\dots,d\} : \hat{\mu}_\alpha > \mu_{\min}\}$. A uniform boud on the error $|\hat{\mu}_\alpha - \mu_\alpha|$ is derived in Goix et al. (2017) which scales roughly as $k^{-1/2}$, where k is the order of magnitude of the number of largest observations used to learn $\mathbb{M}$ and the μ_α's. In Simpson et al. (2018) the original DAMEX framework is refined in order to also model extremes in the directions Ω_α where the angular measure does not concentrate. A third algorithm named CLEF has been proposed by Chiapino and Sabourin (2016) which allows to cluster together different sub-simplices Ω_α's which are close in terms of symmetric difference of the subsets α's. This is particularly useful in situations where the empirical angular mass is scattered onto a large number of sub-simplices, so that DAMEX fails to exhibit a list $\hat{\mathbb{M}}$ of reasonable size. Asymptotic guarantees for the latter approach and variants, leading to statistical tests with controllable asymptotic type I error are derived in Chiapino et al. (2018).

In the present paper, support estimation is only a preliminary step before mixture modeling. We decided to use DAMEX in view of its computational simplicity and the statistical guarantees it offers, considering the fact that its output was very similar to CLEF's on the aeronautics dataset considered in our usecase (see Section 6.2). Using the above mentioned alternatives is certainly possible but for the sake of brevity we shall only present the results obtained using DAMEX as a preprocessing step. We now briefly describe how DAMEX works.

The DAMEX algorithm. Given a dataset $(\mathbf{X}_i)_{i\leq n}$ of independent data distributed as $\mathbf{X}$, DAMEX proceeds as follows. First, replace the unknown marginal distributions F_j with their empirical counterpart $\hat{F}_j(x) = \frac{1}{n}\sum \mathbb{1}\{X_{i,j} < x\}$ and define next $\hat{V}_{i,j} = (1-\hat{F}_j(X_{i,j}))^{-1}$ and $\hat{\mathbf{V}}_i = (\hat{V}_{i,1},\dots,\hat{V}_{i,d})$. Then choose some $k \ll n$ large enough (typically $k = O(\sqrt{n})$) and define $\hat{\mu}_\alpha$ as the empirical counterpart of $\mu(R^\epsilon_\alpha)$ with t replaced by n/k, that is $\hat{\mu}_\alpha = (1/k)\sum_{i=1}^n \mathbb{1}\{\hat{\mathbf{V}}_i \in \frac{n}{k}\mathcal{R}^\epsilon_\alpha\}$. Notice that the above description is a variant of the original algorithm in Goix et al. (2016) which uses thickened cones $\mathcal{C}^\epsilon_\alpha$ instead of $\mathcal{R}^\epsilon_\alpha$. However finite sample guarantees in Goix et al. (2017) are obtained using the latter rather

than the original $\mathcal{C}_\alpha^\epsilon$'s, that is why using the $\mathcal{R}_\alpha^\epsilon$'s is preferred.

3 A Mixture Model For Multivariate Extreme Values

The purpose of this section is to develop a novel mixture model for the angular distribution Φ of the largest instances of the dataset, indexed by $\alpha \in \mathbb{M}$, where $\mathbb{M}$ is Φ's support. Each component $\alpha \in \mathbb{M}$ of the mixture generates instances $\mathbf{V}$ such that V_j is likely to be large for $j \in \alpha$ and the latent variables of the model take their values in $\mathbb{M}$. In practice, we adopt a *plug-in* approach and identify $\mathbb{M}$ with $\widehat{\mathbb{M}}$, the output of DAMEX. As the distribution of extremes may be entirely characterized by the distribution of their angular component $\mathbf{W} \in \mathcal{S}_d$ (see the polar decompositions (2) and (3)), a natural model choice is that of Dirichlet mixtures. We next show how to design a 'noisy' version of the model for subasymptotic observations and how to infer it by means of an EM procedure based on a truncated version of the original dataset, surmounting difficulties related to the geometry of Φ's support.

3.1 Angular Mixture Model for the Directions along which Anomalies Occur

Recall from Section 2.3 that $\mathcal{S}_d$ is naturally partitioned into $2^d - 1$ sub-simplices $\mathcal{S}_\alpha$. Our key assumption is that the support of μ (or Φ) is *sparse* in the sense that $|\mathbb{M}| \ll 2^d$, where $\mathbb{M} = \{\alpha : \mu(\mathcal{C}_\alpha) > 0\} = \{\alpha : \Phi(\mathcal{S}_\alpha) > 0\}$. Let K denote the number of subsets $\alpha \in \mathbb{M}$ of cardinality at least 2 and let $d_1 \in \{0, \ldots, d\}$ be the number of singletons $\{j\} \in \mathbb{M}$. Without loss of generality we assume that these singletons correspond to the first d_1 coordinates, so that $\mathbb{M} = \{\alpha_1, \ldots, \alpha_K, \{1\}, \ldots, \{d_1\}\}$. For simplicity, we also suppose that the sets $\alpha \in \mathbb{M}$ are not nested, an hypothesis which can be relaxed at the price of additional notational complexity. In view of (5), the angular measure then admits the decomposition

$$d^{-1}\Phi(\,\cdot\,) = \sum_{k=1}^{K} \pi_k \Phi_{\alpha_k}(\,\cdot\,) + \sum_{j \le d_1} \pi_{K+j} \delta_{\mathbf{e}_j}(\,\cdot\,),$$

where Φ_{α_k} is a probability measure on $\mathcal{S}_{\alpha_k}$, the weights π_k satisfy $\sum_{k \le K+j} \pi_k = 1$ and $\mathbf{e}_j = (0, \ldots, 1, \ldots, 0)$ is the j^{th} canonical basis vector of $\mathbb{R}^d$. The singletons weights derive immediately from the moment constraint (4): for $i \le d_1$,

$$d^{-1} = \sum_{k=1}^{K} \int_{\mathcal{S}_{\alpha_k}} w_i \pi_k \Phi_{\alpha_k}(\mathrm{d}\mathbf{w}) + \sum_{j \le d_1} \int_{\mathcal{S}_{\{j\}}} w_i \, \pi_{K+j} \delta_{\mathbf{e}_j}(\mathrm{d}\mathbf{w}) = \pi_{K+i}.$$

We obtain

$$\Phi(\,\cdot\,) = d\sum_{k=1}^{K} \pi_k \Phi_{\alpha_k}(\,\cdot\,) + \sum_{j\leq d_1} \delta_{\mathbf{e}_j}(\,\cdot\,), \tag{7}$$

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

$$\sum_{k=1}^{K} \pi_k = 1 - d_1/d. \tag{8}$$

Equation (7) determines the structure of the angular distribution of the largest observations. For likelihood-based inference, a parametric model for each component Φ_{α_k} of the angular measure must be specified. One natural model for probability distributions on a simplex is the Dirichlet family, which provides a widely used prior in Bayesian statistics for data clustering purposes in particular. We recall that the Dirichlet distribution on a simplex $\mathcal{S}_\alpha$ admits a density φ_α with respect to the $(|\alpha|-1)$-dimensional Lebesgue measure which is denoted by d$\mathbf{w}$ for simplicity. It can be parameterized by a mean vector $\mathbf{m}_\alpha \in \mathcal{S}_\alpha$ and a concentration parameter $\nu_\alpha > 0$, so that for $\mathbf{w} \in \mathcal{S}_\alpha$,

$$\varphi_\alpha(\mathbf{w}|\mathbf{m}_\alpha, \nu_\alpha) = \frac{\Gamma(\nu_\alpha)}{\prod_{i\in\alpha}\Gamma(\nu_\alpha m_{\alpha,i})} \prod_{i\in\alpha} w_i^{\nu_\alpha m_{\alpha,i}-1}.$$

Refer to *e.g.* Müller and Quintana (2004) for an account of Dirichlet processes and mixtures of Dirichlet Processes applied to Bayesian nonparametrics. We emphasize that our context is quite different: a Dirichlet Mixture is used here as a model for the angular component of the largest observations, not as a prior on parameters. This modeling strategy for extreme values was first proposed in Boldi and Davison (2007) and revisited in Sabourin and Naveau (2014) to handle the moment constraint (4) via a model re-parametrization. In both cases, the focus was on moderate dimensions. In particular, both cited references worked under the assumption that the angular measure concentrates on the central simplex $\Omega_{\{1,\ldots,d\}}$ only. In this low dimensional context, the main purpose of the cited authors was to derive the posterior predictive angular distribution in a Bayesian framework, using a variable number of mixture components concentrating on $\Omega_{\{1,\ldots,d\}}$. Since the set of Dirichlet mixture distributions with an arbitrary number of components is dense among all probability densities on the simplex, this model permits in theory to approach any angular measure for extremes. The scope of the present paper is different. Indeed we are concerned with high dimensional data (say $d \simeq 100$) and consequently we do not attempt to model the finest details of the angular measure. Instead we intend to design a model accounting only for information which is relevant for clustering. Since an intuitive summary of an extreme event in a high dimensional context is the subset α of features it involves, we assign one mixture component per sub-simplex Ω_α such that $\alpha \in \mathbb{M}$. Thus we model each Φ_α by a single Dirichlet

distribution with unknown parameters m_α, ν_α. Using the standard fact that for such a distribution, $\int_{\mathbb{S}_\alpha} \mathbf{w}\varphi_\alpha(\mathbf{w}|m_\alpha,\nu_\alpha)\mathrm{d}\mathbf{w} = \mathbf{m}_\alpha$, the moment constraint (4) becomes:

$$\frac{1}{d} = \sum_{k=1}^{K} \pi_k \mathbf{m}_{k,j}, \quad j \in \{d_1+1,\ldots,d\}, \tag{9}$$

where $\mathbf{m}_k = \mathbf{m}_{\alpha_k}$ for $k \leq K$.

3.2 A Statistical Model for Large but Sub-asymptotic Observations.

Recall from (3) that Φ is the *limiting* distribution of $\mathbf{V}$ for large R's. In practice, we dispose of no realization of this limit probability measure and the observed angles corresponding to radii $R > r_0$ follow a sub-asymptotic version of Φ. In particular, if the margins V_j have a continuous distribution, we have $\mathbb{P}(V_j \neq 0) = 1$ so that all the $\mathbf{V}_i = (V_{i,1},\ldots,V_{i,d})$, $1 \leq i \leq n$, lie in the central cone $\mathcal{C}_{\{1,\ldots,d\}}$ (this is also true using the empirical versions $\hat{\mathbf{V}}_i$ defined in subsection 2.3). In the approach we propose, the deviation of $\mathbf{V}$ from its asymptotic support, which is $\bigcup_{\alpha\in\mathbb{M}} \mathcal{C}_\alpha$, is accounted for by a noise $\boldsymbol{\varepsilon}$ with light tailed distribution, namely an exponential distribution. That is, we assume that $\mathbf{V} = R\,\mathbf{W} + \boldsymbol{\varepsilon}$, see Model 1 below. As is usual for mixture modeling purposes, we introduce a multinomial latent variable $\mathbf{Z} = (Z_1, \ldots, Z_{K+d_1})$ such that $\sum_k Z_k = 1$ and $Z_k = 1$ if $\mathbf{W}$ has been generated by the k^{th} component of the angular mixture (7). In a nutshell, the type of anomaly/extreme is encoded by the latent vector $\mathbf{Z}$. Then, for $k \leq K$, $\mathbb{P}(Z_k = 1) = \pi_k$, while, for $K < k \leq K + d_1$, $\mathbb{P}(Z_k = 1) = d^{-1}$. The unknown parameters of the model are $\boldsymbol{\theta} = (\pi, \mathbf{m}, \boldsymbol{\nu})$, where $\nu_k > 0$ and $\pi = (\pi_1,\ldots,\pi_K)$, $\mathbf{m} = (\mathbf{m}_1,\ldots,\mathbf{m}_K)$ must satisfy the constraints (8) and (9), as well as the exponential rates $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{K+d_1})$, where $\lambda_k > 0$. Figure 1 illustrates Model 1 in dimension $d = 3$.

Model 1 (Sub-asymptotic mixture model)**.**

- *Consider a standard regularly varying random vector* $\mathbf{V}$ *satisfying* (1) *(typically* $V_j = (1 - \hat{F}_j(X_j))$ *for* $\hat{F}_j$ *an estimate of the marginal distribution* F_j *of* X_j*, see subsection 2.1).*

- *Let* $R = \|\mathbf{V}\|$*. Fix some high radial threshold* r_0*, typically a large quantile of the observed radii. Let* $\mathbf{Z}$ *be a hidden variable indicating the mixture component in* (7)*. Conditionally to* $\{R > r_0, Z_k = 1\}$*,* $\mathbf{V}$ *decomposes as*

$$\mathbf{V} \;=\; \mathbf{V}_k + \boldsymbol{\varepsilon}_k \;=\; R_k \mathbf{W}_k + \boldsymbol{\varepsilon}_k, \tag{10}$$

where $\mathbf{V}_k \in \mathcal{C}_{\alpha_k}$*,* $\boldsymbol{\varepsilon}_k \in \mathcal{C}_{\alpha_k}^{\perp}$*,* $R_k = \|\mathbf{V}_k\|$*,* $\mathbf{W}_k = R_k^{-1}\mathbf{V}_k \in \mathcal{S}_{\alpha_k}$*. The components* $R_k, \mathbf{W}_k, \boldsymbol{\varepsilon}_k$ *are independent from each other. The noise's components are* i.i.d. *according to a translated exponential distribution with rate* λ_k*,* R_k *is Pareto distributed above* r_0 *and* $\mathbf{W}_k$ *is distributed as* Φ_k*, that is*

$$\begin{cases} \mathbb{P}\left(R_k > r\right) = r_0 r^{-1}, r > r_0\,, \\ \mathbf{W}_k \sim \Phi_k\,, \\ \varepsilon_j \sim 1 + \mathcal{E}xp(\lambda_k), j \in \{1, \ldots, d\} \setminus \alpha_k\,, \end{cases}$$

with $\Phi_k = \varphi_k(\,\cdot\,|m_k, \nu_k)$ *if* $k \leq K$*, and* $\Phi_k = \delta_{\mathbf{e}_{k-K}}$ *if* $K < k \leq K + d_1$*.*

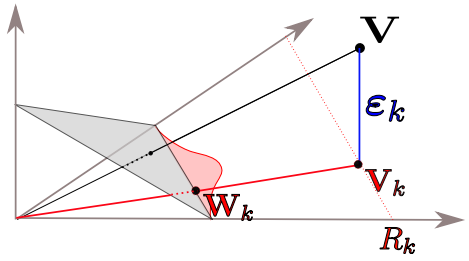

Figure 1: Trivariate illustration of the sub-asymptotic model 1:
the observed point $\mathbf{V}$ has been generated by component $\alpha_k = \{1, 2\}$. The grey triangle is the unit simplex, the shaded red area stands for the Dirichlet density φ_k.

4 Statistical Inference via EM Algorithm.

In the mixture model setting described above with hidden variables Z_i, likelihood optimization is classically performed using an EM algorithm (Dempster et al., 1977). This method consists in performing in turn the so-called E-step

and M-step at each iteration t. Denoting by $\boldsymbol{\theta}_t$ the value at iteration t of the set of unknown model parameters, the posterior probabilities

$$\gamma_{i,k}^{(t+1)} = \mathbb{P}(Z_{i,k} = 1 | \mathbf{V}_i, \boldsymbol{\theta}_t)$$

are computed during the E-step and define the objective function

$$Q(\boldsymbol{\theta}, \gamma^{(t)}) = \sum_i \sum_k \gamma_{i,k}^{(t+1)} \log p(\mathbf{V}_i | Z_{i,k} = 1, \boldsymbol{\theta}).$$

The latter serves as a proxy for the log-likelihood and can be maximized with respect to $\boldsymbol{\theta}$ with standard optimization routines during the M-step, which yields $\boldsymbol{\theta}_{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \gamma^{(t)})$. The procedure stops when the value $Q(\boldsymbol{\theta}_t, \gamma^{(t)})$ reaches a stationary point and the latest pair $(\boldsymbol{\theta}_t, \gamma^{(t)})$ is returned.

The likelihood for Model 1, $p(\mathbf{v}|\boldsymbol{\theta} = (\mathbf{m}, \boldsymbol{\nu}, \pi, \boldsymbol{\lambda}))$, for one observation $\mathbf{v} \in (1, \infty)^d, \|\mathbf{v}\| \geq r_0$, follows directly from the model specification,

$$p(\mathbf{v}|\boldsymbol{\theta}) = r_0 \sum_{k=1}^{K} \pi_k \, r_k^{-|\alpha_k|-1} \, \varphi_k(\mathbf{w}_k | \mathbf{m}_k, \nu_k) \prod_{j \in \alpha_k^c} f_\varepsilon(v_j | \lambda_k) + \frac{r_0}{d} \sum_{k=K+1}^{K+d_1} r_k^{-2} \prod_{j \in \{1,\ldots,d\} \setminus k} f_\varepsilon(v_j | \lambda_k) \quad (11)$$

where $f_\varepsilon(\cdot \,|\lambda_k)$ denotes the marginal density of the noise $\boldsymbol{\varepsilon}_k$ given the noise parameter λ_k. As specified in Model 1, in this paper we set $f_\varepsilon(x|\lambda_k) = \lambda_k e^{-\lambda_k(x-1)}$, $x > 1$ (a translated exponential density), but any other light tailed distribution could be used instead. Notice that the term $r_k^{-|\alpha_k|-1} = r_k^{-2} r_k^{-|\alpha_k|+1}$ is the product of the radial Pareto density and the Jacobian term for the change of variables $T_k : \mathbf{V}_k \mapsto (R_k, \mathbf{W}_k)$. Recall that the constraints are

$$\nu_k > 0 \; (1 \leq k \leq K) \, , \qquad \lambda_k > 0 \; (1 \leq k \leq K + d_1), \quad (12)$$

and that $\pi = (\pi_1, \ldots, \pi_K)$ and $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_K)$ satisfy (8) and (9). The latter linear constraint on $(\boldsymbol{\pi}, \mathbf{m})$ implies that $\mathbf{m}$ and $\boldsymbol{\pi}$ cannot be optimized independently, which complicates the M-step of an EM-algorithm. Thus we begin with a re-parametrization of the model ensuring that the moment constraint (4) is automatically satisfied.

Re-parametrization. In a lower dimensional Bayesian framework, earlier works (Sabourin and Naveau (2014)) have proposed a re-parametrization of the pair $(\boldsymbol{\pi}, \mathbf{m})$ ensuring that the moment constraint (4) is automatically satisfied. This consists in a sequential definition of the mixture centers m_k together with the involving partial barycenters of the remaining components $(m_{k+1}, \ldots, m_k)$. The advantage if this construction is that the resulting parameter has a intuitive interpretation which facilitates the definition of a prior, while allowing for efficient MCMC with reversible jumps sampling (Green (1995)) of the posterior

distribution. However, how to adapt this re-parameterization to our context where several sub-simplices are involved remains an open question and we did not pursue this idea further. The re-parametrization that we propose here consists in working with the product parameter $\rho_{k,j} = \pi_k m_{k,j}$ instead of the pair $(\pi_k, m_{k,j})$. Namely, consider a $K \times (d - d_1)$ matrix $\boldsymbol{\rho} = ({\boldsymbol{\rho}_1}^\top, \ldots, {\boldsymbol{\rho}_K}^\top)$ where $\rho_{k,j} > 0$ for $j \in \alpha_k$ and $\rho_{k,j} = 0$ otherwise. Then, for all $k \in \{1 \ldots, K\}$, set

$$\pi_k := \sum_{j \in \alpha_k} \rho_{k,j} \text{ and } m_{k,j} := \frac{\rho_{k,j}}{\pi_k}, \forall j \in \alpha_k. \tag{13}$$

Then (8) and (9) together are equivalent to

$$\sum_{\{k: j \in \alpha_k\}} \rho_{k,j} = \frac{1}{d}, \quad \forall j \in \{d_1 + 1, \ldots, d\}. \tag{14}$$

In the sequel we denote respectively by $p(\mathbf{v}|\boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\lambda}) := p(\mathbf{v}|\boldsymbol{\pi}, \mathbf{m}, \boldsymbol{\nu}, \boldsymbol{\lambda})$ and $\varphi_k(\mathbf{w}|\boldsymbol{\rho}_k, \nu_k) := \varphi_k(\mathbf{w}|\mathbf{m}_k, \nu_k)$ the likelihood and the Dirichlet densities in the re-parameterized model, where $(\mathbf{m}, \boldsymbol{\pi})$ are obtained from $\boldsymbol{\rho}$ *via* (13). By abuse of notations, let $\boldsymbol{\theta}$ denote in the sequel the set of parameters of the re-parameterized version of Model 1, that is $\boldsymbol{\theta} = (\boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\lambda})$, and let Θ be the parameter space, that is the set of $\boldsymbol{\theta}$'s such that constraints (12) and (14) hold.

EM algorithm. We summarize below the EM algorithm in our framework. Let $n_0 \leq n$ be the number of observations $\mathbf{V}_i$ such that $\|\mathbf{V}_i\| > r_0$. To alleviate notations, we may relabel the indices i so that these observations are $\mathbf{V}_{1:n_0} = (\mathbf{V}_1, \ldots, \mathbf{V}_{n_0})$. Let $\mathbf{Z}_i = (Z_{i,1}, \ldots, Z_{i,K+d_1}), i \leq n_0$ be the hidden variables associated with $\mathbf{V}_{1:n_0}$. Also let $p(\mathbf{v}|\boldsymbol{\theta}, z_k = 1)$ denote the conditional density of $\mathbf{V}$ given $(Z_k = 1, \boldsymbol{\theta})$. In view of the likelihood (11), it is given by

$$p(\mathbf{v}|z_k = 1, \boldsymbol{\theta}) = \begin{cases} r_k^{-|\alpha_k|-1} \, \varphi_k(\mathbf{w_k}|\rho_\mathbf{k}, \nu_k) \, \prod_{j \in \alpha_k^c} f_\varepsilon(v_j|\lambda_k), & (k \leq K) \\ v_k^{-2} \, \prod_{j \in \{1,\ldots,d\} \setminus k} f_\varepsilon(v_j|\lambda_k), & (K < k \leq K + d_1). \end{cases} \tag{15}$$

EM algorithm for Model 1

Input Extreme standardized data $\mathbf{V}_{1:n_0}$.

- **Initialization** Choose a starting value for $\boldsymbol{\theta}$ (See Remark 2).
- **Repeat until convergence**:

  E-step: compute for $1 \leq i \leq n_0$ and $k \leq K+d_1$, $\gamma_{i,k} = \mathbb{P}\left(Z_{i,k} = 1 \mid \mathbf{V}_i, \boldsymbol{\theta}\right)$ according to (17). Set $\boldsymbol{\gamma} = (\gamma_{i,k})_{i \leq n_0, k \leq K+d_1}$.

  M-step: Solve the optimization problem $\max_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta}, \boldsymbol{\gamma})$where $Q(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n_0} \sum_{k=1}^{K+d_1} \gamma_{i,k}\big(\log \pi_k + \log p(\mathbf{V}_i|\boldsymbol{\theta}, z_{i,k} = 1)\big)$ is a lower bound for the likelihood and $\pi_k = \mathbb{P}(Z_{i,k} = 1|\boldsymbol{\theta})$, *i.e.*

$$\pi_k = \begin{cases} \sum_{\ell \in \alpha_k} \rho_{k,l} & \text{for } 1 \leq k \leq K\ , \\ d^{-1} & \text{for } K < k \leq K + d_1\ , \end{cases} \tag{16}$$

  where $p(\mathbf{V}_i|\boldsymbol{\theta}, z_{i,k} = 1)$ is given by (15).Denote by $\boldsymbol{\theta}^\star$ the solution, set $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$.

Remark 2. *In this work the starting values for the concentration parametrers ν_k are set to 20, those for the exponential rates are set to $\lambda_k = 0.01$. Finally, one may easily construct a matrix $\boldsymbol{\rho}$ satisfying the constraint (14) starting with any matrix $\tilde{\boldsymbol{\rho}}$ such that $\tilde{\rho}_{k,j} = 0$ for $j \notin \alpha_k$ and $\tilde{\rho}_{k,j} > 0$ otherwise, and then defining $\boldsymbol{\rho}$ via $\rho_{k,j} = (\sum_{l=1}^{K} \tilde{\rho}_{l,j})^{-1} \tilde{\rho}_{k,j}$.*

We now describe at length the E-step and the M-step of the algorithm.

E-step. The $\gamma_{i,k}$'s are obtained using the Bayes formula, for $1 \leq k \leq K + d_1$,

$$\gamma_{i,k} = p(Z_{i,k} = 1|\mathbf{V}_i, \boldsymbol{\theta}) = \frac{\pi_k\ p(\mathbf{V}_i|z_{i,k} = 1, \boldsymbol{\theta})}{\sum_{\substack{1 \leq \ell \leq K+d_1 \\ \ell \neq k}} \pi_\ell\ p(\mathbf{V}_i|z_{i,\ell} = 1, \boldsymbol{\theta})}, \tag{17}$$

where π_k is defined in (16) and $p(\mathbf{V}_i|Z_{i,k} = 1, \boldsymbol{\theta})$ is given by (15).

M-step. Here optimization of $Q(\boldsymbol{\theta}, \gamma)$ with respect ot $\boldsymbol{\theta} = (\boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\lambda})$ is performed under constraints (12), (14). Since Q decomposes into a function of $(\boldsymbol{\rho}, \boldsymbol{\nu})$ and a function of $\boldsymbol{\lambda}$, and since the constraints on $\boldsymbol{\rho}, \boldsymbol{\nu}$ and $\boldsymbol{\lambda}$ are independent, maximization can be performed independently over the two blocks. Indeed, gathering terms not depending on $\boldsymbol{\theta}$ into a constant C,

$$Q(\boldsymbol{\theta},\boldsymbol{\gamma}) = \sum_{i=1}^{n}\Big[\sum_{k=1}^{K}\gamma_{i,k}\big[\log\pi_k + \log\varphi_k(\mathbf{W}_{i,k}|\boldsymbol{\rho}_k,\nu_k) + \sum_{l\in\alpha_k^c}\log f_\varepsilon(V_{i,l}|\lambda_k)\big] + \sum_{k=K+1}^{K+d_1}\gamma_{i,k}\big[\sum_{\ell\neq k}\log f_\varepsilon(V_{i,l}|\lambda_k)\big]\Big] + C = Q_1(\boldsymbol{\rho},\boldsymbol{\nu}) + Q_2(\boldsymbol{\lambda}) + C,$$

where

$$Q_1(\boldsymbol{\rho},\boldsymbol{\nu}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{i,k}\big[\log\sum_{l\in\alpha_k}\rho_{\mathbf{k}l} + \log\varphi_k(\mathbf{W}_{i,k}|\rho_{\mathbf{k}},\nu_k)\big]$$
$$Q_2(\boldsymbol{\lambda}) = \sum_{i=1}^{n}\sum_{k=1}^{K+d_1}\gamma_{i,k}\sum_{l\in\alpha_k^c}\log f_\varepsilon(V_{i,l}|\lambda_k)\,.$$

Here we set $\alpha_k = \{k-K\}$ for $K < k \leq K+d_1$, in accordance with the notations from Section 3.1. Notice that the dependence of Q_1 and Q_2 on $\boldsymbol{\gamma}$ is omitted for the sake of concision. With these notations

$$\max_{\substack{\boldsymbol{\theta}\text{ s.t.}\\ (12),\ (14)}} Q(\boldsymbol{\theta},\boldsymbol{\gamma}) = \max_{\substack{\boldsymbol{\rho},\boldsymbol{\nu}\text{ s.t.}\\ (14),\nu_k>0,\ k\leq K}} Q_1(\boldsymbol{\rho},\boldsymbol{\nu}) + \max_{\substack{\boldsymbol{\lambda}\text{ s.t.}\\ \lambda_k>0,\ 1\leq k\leq K+d_1}} Q_2(\boldsymbol{\lambda})$$

The function Q_1 being non-concave we use the python package **mystic** (McKerns et al. (2012)) to maximize it. For our choice of translated exponential noise, $f_\varepsilon(v|\lambda_k) = \lambda_k e^{-\lambda_k(v-1)}$, $v \geq 1$, the maximizer of Q_2 has an explicit expression,

$$\lambda_k^* = \frac{|\alpha_k^c|\sum_{i=1}^n\gamma_{i,k}}{\sum_{i=1}^n\gamma_{i,k}\sum_{l\in\alpha_k^c}(V_{i,\ell}-1)}\,, \qquad k \leq K+d_1.$$

Remark 3. *Let $\boldsymbol{\gamma}^t$ and $\boldsymbol{\theta}^t$ be the results of the t-th iteration of the algorithm then we conclude the iterative process if $Q(\boldsymbol{\theta}^t,\boldsymbol{\gamma}^t) < Q(\boldsymbol{\theta}^{t-1},\boldsymbol{\gamma}^{t-1}) + \epsilon$, with ϵ a small threshold.*

5 Graph-based Clustering and Visualization Tools

Beyond the hard clustering that may be straightforwardly deduced from the computation of the likeliest values $z_1, \ldots, z_{n_0}$ for the hidden variables given the $\mathbf{V}_i$'s and the parameter estimates produced by the EM algorithm, the statistical model previously introduced defines a natural structure of undirected weighted graph on the set of observed extremes, which interpretable layouts (graph drawing) can be directly derived using classical solutions. Indeed, a partition (hard clustering) of the set of (standardized) anomalies/extremes $\mathbf{V}_1, \ldots, \mathbf{V}_{n_0}$ is

obtained by assigning membership of each $\mathbf{V}_i$ in a cluster (or cone/sub-simplex) determined by the component of the estimated mixture model from which it arises with highest probability: precisely, one then considers that the abnormal observation $\mathbf{V}_i$ is in the cluster indexed by

$$k_i = \underset{k \in \{1, \ldots, K+d_1\}}{\arg\max} \gamma_{i,k}$$

and is of type α_{k_i}. However, our model-based approach brings much more information and the vector of posterior probabilities $(\gamma_{i,1}, \ldots, \gamma_{i,K+d_1})$ output by the algorithm actually defines soft membership and represent the uncertainty in whether anomaly $\mathbf{V}_i$ is in a certain cluster. It additionally induces a similarity measure between the anomalies: the higher the probability that two extreme values arise from the same component of the mixture model, the more similar they are considered. Hence, consider the undirected graph whose vertices, indexed by $i = 1, \ldots, n_0$, correspond to the extremal observations $\mathbf{V}_1, \ldots, \mathbf{V}_{n_0}$ and whose edgeweights are $w_{\boldsymbol{\theta}}(\mathbf{V}_i, \mathbf{V}_j)$, $1 \leq i \neq j \leq n_0$, where

$$w_{\boldsymbol{\theta}}(\mathbf{V}_i, \mathbf{V}_j) = \mathbb{P}\left(\mathbf{Z}_i = \mathbf{Z}_j \mid \mathbf{V}_i = \mathbf{V}_i,\ \mathbf{V}_j = \mathbf{V}_j,\ \boldsymbol{\theta}\right) = \sum_{k=1}^{K+d_1} \gamma_{i,k}\gamma_{j,k}.$$

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

$$w_{\boldsymbol{\theta}}(\mathbf{V}_i, \mathbf{V}_{(i,1)}) \geq w_{\boldsymbol{\theta}}(\mathbf{V}_i, \mathbf{V}_{(i,2)}) \geq \ldots \geq w_{\boldsymbol{\theta}}(\mathbf{V}_i, \mathbf{V}_{(i,n_0)})$$

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

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

Graph visualization. In possible combination with clustering, graph visualization techniques (see *e.g.* Hu and Shi (2015)), when the number n_0 of anomalies to be analyzed is large, can also be used to produce informative layouts. Discussing the merits and limitations of the wide variety of approaches documented in the literature in this purpose is beyond the scope of this paper. The usefulness of the weighted graph representation proposed above combined with state-of-the-art graph-mining tools is simply illustrated in Section 6.2 and 6.3. We point out however that alternatives to the (force-based) graph drawing method used therein can be naturally considered, re-using for instance the eigenvectors of the graph Laplacian computed through a preliminary spectral clustering procedure (see *e.g.* Athreya et al. (2017) and the references therein for more details on spectral layout methods).

6 Illustrative Experiments

The aim of our experiments is double. First, investigate the goodness of fit of the Dirichlet mixture model fitted *via* the EM algorithm on simulated data from the model. Second, provide empirical evidence of the relevance of the approach we promote for anomaly clustering/visualization with real world data. Comparisons with state-of-the-art methods standing as natural competitors are presented for this purpose.

6.1 Experiments on Simulated Data

To assess the performance of the proposed estimator of the dependence structure and of the EM algorithm, we generate synthetic data according to Model 1. The dimension is fixed to $d = 100$ and the mixture components, that is the elements of $\mathbb{M} = \{\alpha_1, \ldots, \alpha_K\}$, are randomly chosen in the power set of $\{1, \ldots, d\}$ with $K = 50$. The coefficients of the matrix ρ which determines the weights and centers through Eq. (14) in the Supplementary Material is also randomly chosen, then its columns are normalized so that the moment constraint is satisfied. Finally. we fix $\nu_k = 20$ for $1 \leq k \leq K$ and λ_k, $1 \leq k \leq K + d_1$, are successively set to 1, 0.75, 0.5, 0.25 and 0.1 to vary the noise level in the experiments. Then each point $\mathbf{V}_i = R_i \mathbf{W}_i + \boldsymbol{\varepsilon}_i$, $i \leq n$, is generated with probability $\pi_k, k \in \{1, \ldots, K\}$ according to the mixture component $k \leq K$, that is

$$R_i \sim Pareto(1) | \{R_i > r_0\},\ \mathbf{W}_i \sim \Phi_k,\ \varepsilon_{i,j} \sim 1 + \mathcal{E}xp(\lambda_k),\ j \in \{1, \ldots, d\} \setminus \alpha_k,$$

and with probability $\frac{1}{d}$ according to component $k \in \{K, \ldots, K + d_1\}$ in such a way that

$$R_i \sim Pareto(1) | \{R_i > r_0\},\ \mathbf{W}_i = 1,\ \varepsilon_{i,j} \sim 1 + \mathcal{E}xp(\lambda_k),\ j \in \{1, \ldots, d\} \setminus \{k\}.$$

The threshold r_0 above which points are considered as extreme is fixed to 100. On this toy example, the pre-processing step that consists in applying DAMEX for recovering $\mathbb{M}$ produces an exact estimate, so that $\hat{\mathbb{M}} = \mathbb{M}$. Then the procedure described in Algorithm 4 is applied. Tables 1 and 2 show the average absolute errors for the estimates $\widehat{\rho}$, $\widehat{\nu}$ and $\widehat{\lambda}$ on 50 datasets of the n_0 generated extreme points, for $n_0 = 1e+3, 2e+3$, namely

$$\begin{aligned}
\mathrm{err}(\widehat{\rho}) &= \frac{1}{50 \cdot K \cdot d} \sum_{l=1}^{50} \sum_{k=1}^{K} \sum_{j=1}^{d} |\widehat{\rho}_{k,j} - \rho_{k,j}| \\
\mathrm{err}(\widehat{\nu}) &= \frac{1}{50 \cdot K} \sum_{l=1}^{50} \sum_{k=1}^{K} |\widehat{\nu}_k - \nu_k| \\
\mathrm{err}(\widehat{\lambda}) &= \frac{1}{50 \cdot (K + d_1)} \sum_{l=1}^{50} \sum_{k=1}^{K+d_1} |\widehat{\lambda}_k - \lambda_k|
\end{aligned}$$

On this toy example, estimates of the means and weights, as well as those of the noise parameters, are almost exact. In contrast, the estimates of the ν_k's are not that accurate, but, as shown next, this drawback does not jeopardize cluster identification.

Table 1: Average error on the model parameters, $n_0 = 1e3$ extreme points

	$\lambda_k = 1.$	$\lambda_k = 0.75$	$\lambda_k = 0.5$	$\lambda_k = 0.25$	$\lambda_k = 0.1$
$err(\widehat{\rho})$	1.39e-5	1.37e-5	1.57e-5	1.22e-5	2.11e-5
$err(\widehat{\nu})$	5.53	5.81	6.28	6.41	9.06
$err(\widehat{\lambda})$	2.65e-2	2.04e-2	1.19e-2	5.97e-3	3.66e-3

Table 2: Average error on the model parameters, $n_0 = 2e3$ extreme points

	$\lambda_k = 1.$	$\lambda_k = 0.75$	$\lambda_k = 0.5$	$\lambda_k = 0.25$	$\lambda_k = 0.1$
$err(\widehat{\rho})$	9.98e-6	1.12e-5	1.06e-5	1.62e-5	1.64e-5
$err(\widehat{\nu})$	3.23	4.13	4.08	4.29	5.05
$err(\widehat{\lambda})$	1.62e-2	1.2e-2	8.11e-3	4.28e-3	3.11e-3

The performance in terms of cluster identification is measured as follows: for each point $\boldsymbol{v}_i$, the true label $y_i \in \{1, \ldots, K + d_1\}$ is compared with the label obtained *via* assignment to the highest probable component, that is $\hat{y}_i = \arg\max_{k \in \{1, \ldots, K+d_1\}} \gamma_{i,k}$. Table 3 shows the average number of labeling errors for different values of n_0 and λ_k. Figure 2 illustrates the relevance of the

Table 3: Average number of labeling errors

	$\lambda_k = 1.$	$\lambda_k = 0.75$	$\lambda_k = 0.5$	$\lambda_k = 0.25$	$\lambda_k = 0.1$
$n_0 = 1e3$	0.	0.	0.	0.6	264.4
$n_0 = 2e3$	0.	0.	0.4	1.8	537.8

proposed approach regarding anomaly visualization. A test set of size 100 consisting of extreme data is simulated as above, and the corresponding matrix $\widehat{\boldsymbol{\gamma}}$ is computed according to (17) with $\boldsymbol{\theta}$ taken as the output of the training step (*i.e.* Algorithm 4 run with the training dataset of $n_0 = 2e3$ points). Finally an adjacency matrix $w_{\widehat{\boldsymbol{\theta}}}(\mathbf{v}_i, \mathbf{v}_j)$ is obtained as detailed in Section 5, on which we apply spectral clustering in order to group the points according to the similarities measured by w. Graph visualization of w is next performed using the python package 'Networkx' Hagber et al. (2008), that provides a spring layout of the graph according to the Früchtermen-Reingold algorithm, see Früchterman and Reingold (1991). A hard thresholding is applied to the edges in w in order to improve readability: edges (i, j) such that $w_{\widehat{\boldsymbol{\theta}}}(\mathbf{v}_i, \mathbf{v}_j) < \epsilon$ with ϵ a small

threshold are removed. Each cluster output by the spectral clustering method is identified with a specific color.

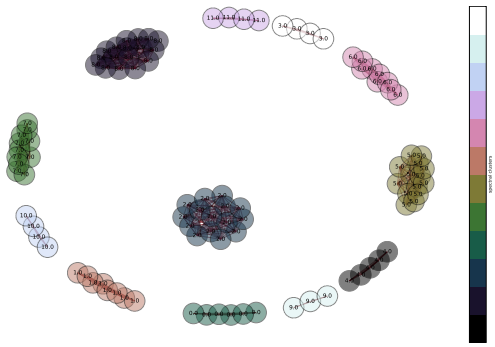

Figure 2: Spectral clustering visualization of a synthetic anomaly test data of size 100 with $d = 20$ and $|\mathbb{M}| = 12$.
Each point is represented as a numbered node. The numbers indicate the true labels, while the colors correspond to the clusters produced by the spectral clustering method. The spatial arrangement of the nodes is obtained by the Früchtermen-Reingold algorithm.

6.2 Flights Clustering and Visualization

The methodology proposed is currently tested by Airbus to assist in building health indicators for condition based maintenance. Health indicators are used for assessing the current state of some system and also for forecasting its future states and possible degradation (*e.g.* bleed, power systems, engine, APU, ...). Airlines can be then informed that some systems should be maintained, so as to avoid any operational procedure at a given time horizon susceptible to cause *e.g.* delays, operational interruptions, *etc* .... The construction of a health indicator can be basically summarized as follows:

1. Collect health and usage data from various aircrafts (generally one has to consider similar ones).
2. Collect some operational events happening on these aircrafts due to some aircraft system errors (*e.g.* operational interruption, delays)
3. Identify anomalies in the health and usage data.
4. Identify some dependencies between health and usage data anomalies and operational events (by means of statistical hypothesis testing but also thanks to human expertise).
5. As soon as some dependencies are well identified, a health indicator is built.

The main barrier is the identification and the understanding of the anomalies. Different operational events are often recorded, corresponding to the degradation of different systems. Usually, a first stage of anomaly detection is performed, followed by a clustering of the anomalies listed for interpretation purpose. The major advantage of the approach proposed in this paper is that it directly provides a similarity measure between the anomalies. This strategy is illustrated by Fig. 3. The proposed method was applied on a dataset of 18553 flights, each of which is characterized by 82 parameters. In order to differentiate between anomalies corresponding to unusually large and small values, each feature is duplicated and each copy of a given feature is defined as the positive (*resp.* negative) value of the parameter above (*resp.* below) its mean value.

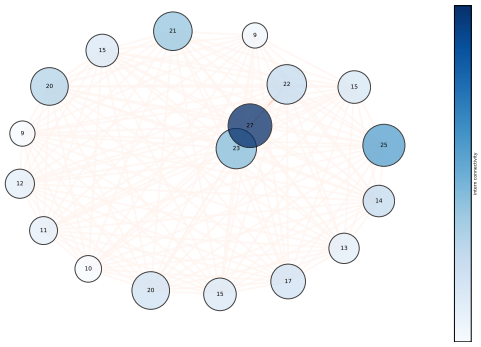

Figure 3: Spectral clustering visualization of flights anomalies with agglomerated nodes.
The agglomerated visualization is obtained *via* spectral clustering: each node represents a cluster. Levels of blue show the intern connectivity between the original nodes so that darker clusters have strongly connected elements. The size of each node is proportional to the number of points forming the cluster.

Fig. 3 and Fig. 4 display the clustering of 300 'extremal' flights into 18 groups, showing on the one hand the output of the spectral clustering applied to the similarity graph $w_{\hat{\boldsymbol{\theta}}}$ and on the other hand the underlying graph obtained with the same procedure as in Fig. 2.

6.3 A Real World Data Experiment with the Ground Truth

The *shuttle* dataset is available in the UCI repository, see Dheeru and Karra Taniskidou (2017) (training and test datasets are merged here), 9 numerical attributes and 7 classes are observed. Class 1 representing more than 80% of the dataset, since our goal is to cluster rare and extreme events, instances from all classes but 1 are analyzed, leading to a sample size equal to 12414. The number of extreme points considered is denoted by n_0 here. We compare our approach to the K-means algorithm and the *spectral clustering* algorithm as implemented in Pedregosa et al. (2011). The number of clusters that we fix in advance to

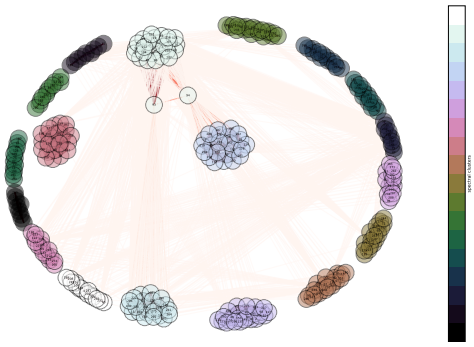

Figure 4: Spectral clustering visualization of flights anomalies.
The number of each node is the (anonymized) flight identification number. The nodes colors and the spatial arrangement are obtained similarly to Fig. 2.

run each of these algorithms is denoted by $n_{cluster}$. The performance of each approach is evaluated by computing the *purity* score:

$$purity = \frac{1}{n_0} \sum_{i=1}^{n_{cluster}} \max_{c \in C} n_{i,c},$$

where $n_{i,c}$ is the number of elements of class c in the cluster i. As shown by Table 4, the purity score produced by the anomaly clustering technique promoted in this paper is always equal to or higher than those obtained by means of the other algorithms.

Table 4: Purity score - Comparisons with standard approaches for different extreme sample sizes.

	$n_0 = 500$	$n_0 = 400$	$n_0 = 300$	$n_0 = 200$	$n_0 = 100$
Dirichlet mixture	0.8	0.82	0.82	0.84	0.85
Kmeans	0.72	0.73	0.75	0.78	0.8
Spectral clustering	0.78	0.77	0.82	0.81	0.8

7 Conclusion

Because extreme values (viewed as anomalies here) cannot be summarized by simple meaningful summary statistics such as local means or modes/centroids, clustering and dimensionality reduction techniques for such abnormal observations must be of very different nature than those developed for analyzing data lying in high probability regions. This paper is a first attempt to design a

methodology fully dedicated to the clustering and visualization of anomalies, by means of a statistical mixture model for multivariate extremes that can be interpreted as a noisy version of the angular measure, which distribution on the unit sphere exhaustively describes the limit dependence structure of the extremes. Mixture component are identified here with specific sub-simplices forming the support of the angular measure. Considering synthetic and real datasets, we also provide empirical evidence of the usefulness of (graph-based) techniques that can be straightforwardly implemented from the framework we developed.

Acknowledgements

This work was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH.

References

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

Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. Wiley.

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

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

Chiapino, M. and Sabourin, A. (2016). Feature clustering for extreme events analysis, with application to extreme stream-flow data. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 132–147. Springer.

Chiapino, M., Sabourin, A., and Segers, J. (2018). Identifying groups of variables with the potential of being large simultaneously. *arXiv preprint arXiv:1802.09977*.

Clifton, D., Hugueny, S., and Tarassenko, L. (2011). Novelty detection with multivariate extreme value statistics. *J. Sign. Proc. Syst.*, 65(3):371–389.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag, London.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *JRSS, Series B (methodological)*, pages 1–38.

Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.

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

Früchterman, T. and Reingold, E. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164.

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

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

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

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

Gorinevsky, D., Matthews, B., and Martin, R. (2012). Aircraft anomaly detection using performance models trained on fleet data. In *Proceedings of the 2012 Conference on Intelligent Data Understanding*.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Hagber, A., Schult, D., and Swart, P. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA.

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

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

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

Liu, F., Ting, K., and Zhou, Z. (2008). Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 413–422.

McKerns, M., Strand, L., Sullivan, T., Fang, A., and Aivazis, M. (2012). Building a framework for predictive science. *arXiv preprint arXiv:1202.1056*.

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

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

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

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

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

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

Roberts, S. (2000). Extreme value statistics for novelty detection in biomedical signal processing. In *Advances in Medical Signal and Information Processing, 2000*, pages 166–172.

Sabourin, A. and Naveau, P. (2014). Bayesian dirichlet mixture model for multivariate extremes: a re-parametrization. *Computational Statistics & Data Analysis*, 71:542–567.

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

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., and Williamson, R. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Simpson, E., Wadsworth, J., and Tawn, J. (2018). Determining the dependence structure of multivariate extremes. *arXiv preprint arXiv:1809.01606*.

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

T. Fawcett, F. P. (1997). Adaptive fraud detection. *Data-Mining and Knowledge Discovery*, 1:291–316.

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

Viswanathan, K., Choudur, L., Talwar, V., Wang, C., Macdonald, G., and Satterfield, W. (2012). Ranking anomalies in data centers. In R.D.James, editor, *Network Operations and System Management*, pages 79–87. IEEE.