Feature Clustering for Support Identification in Extreme Regions

Hamid Jalalzai, Rémi Leluc

Feature Clustering for Support Identification in Extreme Regions

Hamid Jalalzai [1 2] **Rémi Leluc** [1]

Abstract

Understanding the complex structure of multivariate extremes is a major challenge in various fields from portfolio monitoring and environmental risk management to insurance. In the framework of multivariate Extreme Value Theory, a common characterization of extremes' dependence structure is the angular measure. It is a suitable measure to work in extreme regions as it provides meaningful insights concerning the subregions where extremes tend to concentrate their mass. The present paper develops a novel optimization-based approach to assess the dependence structure of extremes. The support identification of extremes scheme rewrites as estimating *clusters of features* which best capture the support of extremes. The dimension reduction technique we provide is applied to statistical learning tasks such as feature clustering and anomaly detection. Numerical experiments provide strong empirical evidence of the relevance of our approach.

1. Introduction

In a wide variety of applications ranging from structural engineering to finance, *extreme* events can occur with a far from negligible probability (Embrechts et al., 1999; 2013). In the multivariate setting, such events are usually modeled through threshold exceedance. A random vector $X = (X^1, \ldots, X^p) \in \mathbb{R}^p, (p > 1)$ is said to be extreme if $\|X\| > t$ for any given norm $\|\cdot\|$ and some *large* threshold $t > 0$. The latter is generally chosen so that a small but non negligible proportion of data falls in the extreme regions $\{x \in \mathbb{R}^p, \|x\| > t\}$. In machine learning tasks, it is relevant to apply different treatments to *extreme* and *normal* data. Devoting attention to extreme regions can lead to better understanding of the distributional law of X and practical performance of classical algorithms, as shown by several recent studies: in anomaly detection (Roberts, 1999; Clifton et al., 2011; Goix et al., 2016; Thomas et al., 2017), classification (Vignotto & Engelke, 2018; Jalalzai et al., 2018; 2020) or feature clustering (Chautru, 2015; Chiapino et al., 2019; Janßen et al., 2020) when dedicated to the most extreme regions of the sample space.

Scaling up multivariate Extreme Value Theory (EVT) is a key issue when addressing high-dimensional learning tasks. Indeed, most multivariate extreme value models have been designed to handle moderate dimensional problems, *e.g.*, where dimension $p \leq 10$. For larger dimensions, simplifying modeling choices are required, stipulating for instance that only some predefined subgroups of components may be concomitant extremes, or, on the contrary, that all must be (Stephenson, 2009; Sabourin & Naveau, 2014).This calls for dimensionality reduction devices adapted to multivariate extreme values.

Identifying the features X^j's (and the resulting subspaces) contributing to X being extreme is a major challenge in EVT. The distributional structure of extremes highlights the components of a multivariate random variable that may be simultaneously large while the others remain small. This is a valuable piece of information for multi-factor risk assessment or detection of anomalies among other –not abnormal– extreme data. Two phenomena are likely to happen: *(i)* only a small number of features may be concomitantly large, so that only a small number of subspaces have non-zero mass, *(ii)* each of these groups -*clusters of features*- contains a limited number of coordinates (compared to the original dimensionality), so that the corresponding subspace with non zero mass have small dimension compared to p. The purpose of this paper is to introduce a data-driven methodology for identifying such subspaces, to reduce the dimensionality of the problem and thus to learn a sparse representation of extreme behaviors.

This paper provides a novel optimization approach to find subspaces from multivariate extreme features. Given $n \geq 1$ *i.i.d* copies $X_1, \ldots, X_n$ of a heavy-tailed random variable $X = (X^1, \ldots, X^p) \in \mathbb{R}^p$, the goal is to identify clusters of features $K \subset [\![1, p]\!]$ such that the variables $\{X^j : j \in K\}$ may be large while the other variables X^j for $j \notin K$ simultaneously remain small. Figure 1 depicts an example of normalized extremes with the associated feature clusters.

[1]Télécom Paris, Institut Polytechnique de Paris, France [2]INRIA, Institut Polytechnique de Paris, France. Correspondence to: Hamid Jalalzai <hamid.jalalzai@inria.fr>, Rémi Leluc <remi.leluc@telecom-paris.fr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

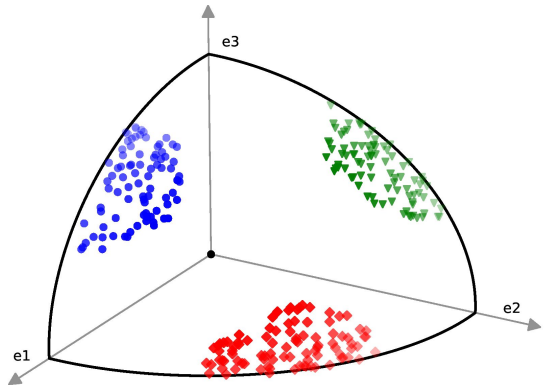

Figure 1. Illustration of normalized extremes θ_i's on the ℓ_2-sphere of $\mathbb{R}^3_+$ with clusters of features $K_1 = \{1,3\}$(blue), $K_2 = \{1,2\}$(red) and $K_3 = \{2,3\}$(green).

Up to approximately 2^p combinations of extreme features are possible and contributions such as Chautru (2015); Chiapino & Sabourin (2016); Goix et al. (2016); Engelke & Hitz (2018); Chiapino et al. (2019) tend to identify a smaller number of simultaneous extreme features. Dimensional reduction methods such as principal components analysis and derivatives (Wold et al., 1987; Cutler & Breiman, 1994; Tipping & Bishop, 1999; Cooley & Thibaud, 2019; Drees & Sabourin, 2019) can be designed to find a lower dimensional subspace where extremes tend to concentrate. Following this path, the idea of the present paper is to decompose the ℓ_1-norm of a positive input sample as a weighted sum of its features.

Several EVT contributions are aimed at assessing a sparse support of multivariate extremes (De Haan & Ferreira, 2007; Chiapino & Sabourin, 2016; Meyer & Wintenberger, 2019; Engelke & Ivanovs, 2020). A broader scope of contributions related to the work detailed in this paper ranges from compressed sensing (Candès et al., 2006; Candes et al., 2006; Tsaig & Donoho, 2006) and matrix factorization (Lee & Seung, 2001; Şimşekli et al., 2015) to group sparsity (Yuan & Lin, 2006; Simon et al., 2013; Devijver et al., 2015).

Contributions. The main results of this paper are:
(i) We present a novel optimization-based approach to perform subspace clustering of extreme regions in the multivariate framework. This is achieved by the algorithm *Multivariate EXtreme Informative Clustering by Optimization* (in short MEXICO) which finds a sparse representation for the dependence structure of extremes.
(ii) Following contribution laid out by Niculae et al. (2018), we study at length different manifolds on the probability simplex, including our $\mathbb{M}$-*set*. Our analysis may be of independent relevance.
(iii) The performance of the introduced algorithm are demonstrated from both theoretical and empirical points of view. First we provide a non-asymptotic bound on the excess risk. Secondly, numerical experiments on both *feature clustering* and *anomaly detection* tasks in extreme regions demonstrate the relevance of our method when compared to existing methods.

Notations. The following notations are used throughout the paper: $\mathcal{M}_n^p([1,+\infty[)$ is the set of $n \times p$ matrices valued in $[1,+\infty[$. Any matrix is denoted in bold. $\mathcal{A}_p^m$ denotes the set of *mixture matrices* composed of $p \times m$ matrices valued in $[0,1]$ where the sum of elements of any column equals 1. For any $\mathbf{M} = (\mathbf{M}_i^j) \in \mathcal{M}_n^p(\mathbb{R})$, for $i \in [\![1,n]\!]$ (*resp.* $j \in [\![1,p]\!]$), let e_i (*resp.* e^j) denote the vector of the canonical basis such that $e_i\mathbf{M} = \mathbf{M}_i$ (*resp.* $\mathbf{M}e^j = \mathbf{M}^j$) where M_i corresponds to the i-th line of $\mathbf{M}$ (*resp.* M^j corresponds to the j-th column). Denote $\mathfrak{S}_m$ the finite symmetric group of order m. Let $E = [0,\infty]^p \backslash \{0\}$ and $\Omega_{p,||\cdot||} = \{x \in \mathbb{R}_+^p : \|x\| \leq 1\}$ the ball associated to the norm $||\cdot||$ and its complementary set $\Omega^c_{p,||\cdot||} = \mathbb{R}_+^p \backslash \Omega_p$, let S denote the sphere associated to $\|\cdot\|$ and for $x \in \mathbb{R}^p$ and $K \subset [\![1,p]\!]$, write $x^{(K)} = (x^j \mathbb{1}_{j \in K})$. Denote by Γ the Euler function.

Outline. The paper is organized as follows, in Section 2 we introduce the multivariate EVT background and our problem of interest. In Section 3 we present our optimization-based approach along with its specific details concerning the projection step onto the probability simplex. Section 4 gathers the theoretical results. We perform some numerical experiments in Section 5 to highlight the performance of our method and we finally conclude in Section 6. Proofs, technical details and additional results can be found in the supplementary material.

2. Preliminaries

Extreme value theory develops models for learning the unusual rather than the usual, in order to provide a reasonable assessment of the probability of occurrence of *rare events*. This section first recalls the required mathematical framework and classical tools for the analysis of multivariate extremes and then introduce our problem of interest.

2.1. Mathematical background

The notion of *regular variation* is a natural way for modelling power law behaviors that appear in various fields of probability theory. In this paper, we shall focus on the dependence and regular variation of random variables and random vectors. We refer to the book of Resnick (1987) for an excellent account of heavy-tailed distributions and the theory of regularly varying functions.

Definition 1 *(Regular variation (Karamata, 1933)) A positive measurable function g is regularly varying with index $\alpha \in \mathbb{R}$, notation $g \in \mathcal{R}_\alpha$ if $\lim_{x \to +\infty} g(tx)/g(x) = t^\alpha$ for all $t > 0$.*

The notion of regular variation is defined for a random variable X when the function of interest is the distribution tail of X.

Definition 2 *(Univariate regular variation) A non-negative random variable X is regularly varying with tail index $\alpha \geq 0$ if its right distribution tail $x \mapsto \mathbb{P}(X > x)$ is regularly varying with index $-\alpha$, i.e., $\lim_{x\to+\infty} \mathbb{P}(X > tx \mid X > x) = t^{-\alpha}$ for all $t > 1$.*

This power-law behavior may be thought of as a smoothness condition for the tail at infinity. This definition can be extended to the multivariate setting where the topology of the probability space is involved. We rely on the vague convergence of measures (Resnick, 1987, Section 3.4) and consider the following definition (Resnick, 1986, p.69).

Definition 3 *(Multivariate regular variation) A random vector $X \in \mathbb{R}^p_+$ is regularly varying with tail index $\alpha \geq 0$ if there exists $g \in \mathcal{R}_{-\alpha}$ and a nonzero Radon measure μ on E such that*

$$g(t)^{-1}\mathbb{P}\left(t^{-1}X \in A\right) \xrightarrow[t\to\infty]{} \mu(A),$$

where $A \subset E$ is any Borel set such that $0 \notin \partial A$ and $\mu(\partial A) = 0$.

The limiting measure μ, known as the *exponent measure*, is homogeneous of order $-\alpha$ *i.e.* for any $t > 0$, $\mu(t\cdot) = t^{-\alpha}\mu(\cdot)$. This suggests a polar decomposition of μ into a radial component and an angular component Φ. For any $x = (x_1, \cdots, x_p) \in \mathbb{R}^p_+$, one can set

$$\begin{cases} R(x) & = ||x|| \\ \Theta(x) & = \left(\frac{x_1}{R(x)}, \ldots, \frac{x_p}{R(x)}\right) \in S \end{cases}$$

For any $B \subset S$, the *angular measure* Φ on S is defined as,

$$\Phi(B) \overset{\text{def}}{=} \mu(\{x, R(x) \geq 1, \Theta(x) \in B\}).$$

The angular measure Φ plays a central role in the analysis of extremes, as it characterizes the directions where extremes are more likely to occur. Assessing the support of Φ, or equivalently of μ, leads to forecasting the directions where extremes are more likely to occur *i.e.* features that are more likely to jointly be large.

2.2. Probabilistic Framework & Problem Statement

We observe $n \geq 1$ *i.i.d* copies $X_1, \ldots, X_n$ of a regularly varying random vector $X = (X^1, \ldots, X^p) \in \mathbb{R}^p$ with tail index $\alpha = 1$. Extremes correspond to samples with norm larger than a fixed threshold $t > 0$. Incidentally, t should depend on n, as the notion of *extreme* should be understood as *large* norms compared to the vast majority of observed data. The Euclidian space $\mathbb{R}^p$ being of finite dimension, all norms are equivalent and the choice of the norm does not matter for the definition of the limit measure (Beirlant et al., 2006), therefore we may use the ℓ_∞-norm in the remainder of this paper to analyse extremes. In other words $R(x)$ is set as $||.||_\infty$ in Definition 3. The observations are first sorted by decreasing order of magnitude $\|X_{(1)}\|_\infty \geq \ldots \geq \|X_{(n)}\|_\infty$. Then, consider a small fraction $0 < \gamma < 1$ of the observations and denote by t_γ the quantile of $||X||_\infty$ at level $(1-\gamma)$, *i.e.* $\mathbb{P}(||X||_\infty > t_\gamma) = \gamma$. The extreme samples are $X_{(1)}, \ldots, X_{(k)}$ where $k = \lfloor n\gamma \rfloor$ is a discrete selection threshold induced by γ (cf Remark 2).

Remark 1 *(Pareto Standardization) In this work, it is assumed that all marginal distributions are tail equivalent to the Pareto distribution with index $\alpha = 1$. In practice, the tails of the marginals may be different and it is convenient to work with marginally standardized variables. Thus, the margins $F^j(x^j) = \mathbb{P}\left(X^j \leq x^j\right)$ are separated from the dependence structure in the description of the joint distribution of X. Consider the Pareto standardized variables $V^j = 1/(1 - F^j(X^j)) \in [1, \infty]$ and $V = T(X) = (V^1, \ldots, V^d)$. Replacing X by V permits to take $\alpha = 1$ and $g(t) = 1/t$ in Definition 3. Appendix B.1 provides further details concerning $\widehat{T}$ the empirical counterpart of T.*

Remark 2 *(On selection of k) Determining k is a central bias variance trade-off of Extreme Value analysis (See e.g. Goix et al. (2016) and references therein). As k gets too large, a bias is induced by taking into account observations which do not necessarily behave as extremes: their distribution deviates significantly from the limit distribution of extremes. On the other hand, too small values lead to an increase of the algorithm's variance.*

Our work focuses on assessing the dependence structure in extreme regions in a multivariate setup. The angular measure Φ fully describes the latter *asymptotic* dependence. Therefore, we seek to accurately infer a *sparse* summary of the mass of extremes spread on each constructed subspace. Let $X \in \mathbb{R}^p_+$ be a multivariate random vector whose dependence structure is unknown. We address the problem of finding different feature clusters $K_j \subset [\![1, p]\!]$ with $j = 1, \ldots, m$ and $m < p$ such that all features in a same subset may be large together. In order to reach a representation of interest, *e.g.*, diversity for portfolio in finance or clusters for smart grids in wireless technologies, we seek disjoint clusters ($K_i \cap K_j = \emptyset$ for all $i \neq j$). Relying on the m clusters of features $K_1, \ldots, K_m$ and the underlying subspaces, the exponent measure μ can be approximated as $\mu(\cdot) \approx \sum_{j=1}^m \mu_{K_j}(\cdot)$. Each component μ_{K_j} is concentrated on the subregion given by the features of cluster K_j.

In the remaining of this paper, $\mathbf{X} \in \mathcal{M}_{k,p}([1,+\infty[)$ corresponds to the truncated training set: $X_{(1)}, \ldots, X_{(k)}$. We search a subset K of features such that the ℓ_1-norms of $\mathbf{X}_i$ and its restriction $\widetilde{\mathbf{X}}_i = \mathbf{X}_i^{(K)}$ are almost equal *i.e.*

$$\|\widetilde{\mathbf{X}}_i\|_1 \approx \|\mathbf{X}_i\|_1.$$

3. Feature Mixture in Extreme Regions

This section presents an approach to find relevant directions of the extreme samples to estimate the support of μ. The analysis is carried out under the empirical risk minimization paradigm and details our algorithm MEXICO.

3.1. Empirical Risk Minimization

To assess the dependence structure of features of extreme samples, we consider the framework of empirical risk minimization focused on extreme regions. Consider an extreme sample X, *i.e.*, an observation satisfying $\|X\|_\infty > t_\gamma$. The goal is to learn a representation function $h : \mathbb{R}^p \to \mathbb{R}_+$ in order to minimize the Bayes risk at level t_γ defined by

$$\mathcal{R}_{t_\gamma}(h) = \mathbb{E}_X \left[\ell\left(X, h\left(X\right)\right) \middle| \|X\|_\infty > t_\gamma \right], \qquad (1)$$

where $\ell : \mathbb{R}_+^p \times \mathbb{R}_+ \to \mathbb{R}_+$ is a loss function measuring the discrepancy between the true extreme dependence structure of X and its predicted counterpart $h\left(X\right)$. Based on the extreme observations $X_{(1)}, \ldots, X_{(k)}$, the empirical risk in the extreme regions is given by

$$\widehat{\mathcal{R}}_k(h) = \frac{1}{k} \sum_{i=1}^{k} \ell\big(X_{(i)}, h(X_{(i)})\big). \qquad (2)$$

Features mixtures. In order to recover the clusters of features, we consider mixtures of the components of each sample. The true number of subregions is unknown and we search for m clusters where m is selected according to Remark 3. We consider the probability simplex Δ_p defined on the positive orthant of $\mathbb{R}_+^p$ by

$$\Delta_p = \{x \in \mathbb{R}_+^p, x_1 + \ldots + x_p = 1\},$$

and let $\mathbf{W} \in \mathcal{A}_p^m$ with $m < p$ be a *mixture matrix*. We denote by $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{W} \in \mathcal{M}_k^m(\mathbb{R}_+)$ the transformed matrix. The following proposition ensures the preservation of the regular variation of the resulting vectors $\widetilde{X}_i$'s.

Proposition 1 (Mixture transformation) *Let $X \in \mathbb{R}_+^p$ be a regularly varying vector as defined earlier and $\mathbf{W} \in \mathcal{A}_p^m$ a mixture matrix with $1 < m \leq p$. Then the transformed vector $\widetilde{X} = X\mathbf{W} \in \mathbb{R}_+^m$ is regularly varying with tail index $\alpha = 1$. Thereby, if we denote by μ (resp. $\widetilde{\mu}$) the limiting measure of X (resp. $\widetilde{X}$), we have*

$$(1/m)\widetilde{\mu}(\Omega^c_{m,||\cdot||_1}) \leq \widetilde{\mu}(\Omega^c_{m,||\cdot||_\infty}) \leq \mu(\Omega^c_{p,||.||_1}).$$

The proof of the proposition is deferred to the Supplementary Material.

Remark 3 *(Selection of m) In view of Proposition 1, in practice, the required dimension $m < p$ can be seen as the smallest value m such that the empirical version of $\widetilde{\mu}(\Omega^c_{m,||.||_\infty})$ is arbitrarly close to the empirical version of $\mu(\Omega^c_{p,||.||_1})$. Hence, the m selected clusters provide relevant support of extremes.*

Loss function. A natural question rises in the choice of the approximation function g used in Eq. (1). Each column $\mathbf{W}^j$ for $j \in [\![1, m]\!]$ is modelling a mixture of components and represents a cluster K_j. For any sample X, we want to find a mixture that gives a good approximation in ℓ_1-norm, *i.e.*, we seek a column $j \in [\![1, m]\!]$ for which $\widetilde{X}^j = (X\mathbf{W})^j$ is the closest to $\|X\|_1$. A simple choice for the approximation function h is reached through a linear combination and defined as follows for any input $x \in \mathbb{R}_+^p$,

$$h_{\mathbf{W}} : x \mapsto h_{\mathbf{W}}(x) = \max_{1 \leq j \leq m} (x\mathbf{W})^j. \qquad (3)$$

The associated loss function is defined by

$$\ell\left(x, h_{\mathbf{W}}(x)\right) = \frac{1}{p}\left(\|x\|_1 - h_{\mathbf{W}}(x)\right). \qquad (4)$$

Observe that this particular choice yields a loss function bounded in $[0, 1]$ when using the angular decomposition of extremes $\theta = X/\|X\|_\infty$. With this choice, the approximation function $h_{\mathbf{W}}$ is parametrized by the mixture matrix $\mathbf{W}$. For ease of notation we abusively denote by $\ell(X, \mathbf{W}) = \ell(X, h_{\mathbf{W}}(X))$. Thus, using this specific loss in Eq. (1), the goal is to learn the mixture matrix $\mathbf{W}_t^\star$ minimizing the risk on extremes

$$\mathbf{W}_t^\star \in \operatorname*{arg\,min}_{\mathbf{W} \in \mathcal{A}_p^m} \left\{ \mathcal{R}_t(\mathbf{W}) \overset{\text{def}}{=} \mathbb{E}\left[\ell\left(X, \mathbf{W}\right) | \|X\|_\infty > t\right] \right\}.$$

Based on the observations $\{X_{(1)}, \ldots, X_{(k)}\}$, the optimization problem consists in finding a mixture matrix minimizing the empirical risk

$$\widehat{\mathbf{W}}_k \in \operatorname*{arg\,min}_{\mathbf{W} \in \mathcal{A}_p^m} \left\{ \widehat{\mathcal{R}}_k(\mathbf{W}) = \frac{1}{k} \sum_{i=1}^{k} \ell\left(X_{(i)}, \mathbf{W}\right) \right\} \qquad (5)$$

Note that $\mathcal{A}_p^m$ is a closed and bounded set hence compact (Bourbaki, 2007) thus there exists at least one solution which can be reached. The minimization problem of Eq. (5) can be rewritten as

$$\widehat{\mathbf{W}}_k \in \operatorname*{arg\,max}_{\mathbf{W} \in \mathcal{A}_p^m} \frac{1}{k} \sum_{i=1}^{k} h_{\mathbf{W}}(X_{(i)}).$$

The index of the column representing a good mixture can be defined with the mapping

$$\varphi : [\![1, k]\!] \to [\![1, m]\!], \quad \varphi(i) = \operatorname*{arg\,max}_{1 \leq j \leq m} \widetilde{X}^j_{(i)}$$

and the optimization problem becomes

$$\underset{\mathbf{W}\in\mathcal{A}_p^m}{\arg\max}\left\{\frac{1}{k}\sum_{i=1}^{k}(\mathbf{XW})_i^{\varphi(i)}=\frac{1}{k}\sum_{i=1}^{k}e_i(\mathbf{XW})e^{\varphi(i)}\right\}. \tag{6}$$

Illustrative example. As a first go, consider the following example showing the way the matrix $\mathbf{W}$ recovers the different clusters. Assume that the vector $X\in\mathbb{R}_+^p$ is exactly coming from a mixture of m disjoint clusters $K_1,\dots,K_m$ and for each sample $\mathbf{X}_i$, there exists K_j such that $\|\mathbf{X}_i^{(K_j)}\|_1=\|\mathbf{X}_i\|_1$. For all $j\in[\![1,m]\!]$, denote $U^j\in[0,1]^p$ the uniform vector with support K_j, *i.e.*, $U^j=(1/|K_j|)^{(K_j)}$. A solution to the optimization problem is given by any column-permutation of the matrix $\mathbf{W}$ whose columns are the vectors U^j. Indeed, the transformed data matrix is $\widetilde{\mathbf{X}}=\mathbf{XW}$ and for any sample $\mathbf{X}_i$ whose features are coming from a cluster K_j, we have

$$\forall l\neq j,\quad \widetilde{\mathbf{X}}_i^j=\mathbf{X}_iU^j=\mathbf{X}_i^{(K_j)}U^j\geq\mathbf{X}_iU^l=\widetilde{\mathbf{X}}_i^l.$$

Taking $\varphi(i)=\arg\max_{1\leq l\leq m}\widetilde{\mathbf{X}}_i^l$ exactly recovers the cluster of index $j=\varphi(i)$. In the case where the large features of the different sample $\mathbf{X}_i$ are all equal, then the columns of the mixture matrix $\mathbf{W}$ tend exactly to uniform vectors with restricted support.

3.2. Optimization on the Simplex

Problem relaxation. One can directly solve the linear program (6) but this formulation suffers from drawbacks. First, the solution could belong to a vertex of the simplex and would induce a unique direction. Second, it involves finding the mapping φ among all the possible combinations which can be prohibited when k or p increases. Thus, one can solve a relaxed version of Eq. (6) by introducing another matrix of mixtures $\mathbf{Z}\in\mathcal{A}_m^k$. The relaxed problem is

$$(\widehat{\mathbf{W}}_k,\widehat{\mathbf{Z}}_k)\in\underset{(\mathbf{W},\mathbf{Z})\in\mathcal{A}_p^m\times\mathcal{A}_m^k}{\arg\max}\frac{1}{k}\sum_{i=1}^{k}\mathbf{X}_i\mathbf{WZ}^i. \tag{7}$$

Optimization problem. We recognize the trace operator in Eq. (7) which is linear and can define an objective function $f:\mathcal{A}_p^m\times\mathcal{A}_m^k\to\mathbb{R}$ that we need to maximize:

$$\begin{cases}(\widehat{\mathbf{W}}_k,\widehat{\mathbf{Z}}_k)\in\arg\max_{(\mathbf{W},\mathbf{Z})}f(\mathbf{W},\mathbf{Z})\\ f(\mathbf{W},\mathbf{Z})=Tr(\mathbf{XWZ})/k\end{cases}$$

The objective function f is bilinear in finite dimension hence continuous. Since maximization occurs on compact sets, there is at least one solution $(\widehat{\mathbf{W}}_k,\widehat{\mathbf{Z}}_k)$. However, it is not unique since any column-permutation of $\widehat{\mathbf{W}}_k$ along with the associated row-permutation of $\widehat{\mathbf{Z}}_k$ is also a valid solution. Indeed, any column (*resp.* row) permutation consists of a multiplication on the right (*resp.* left) side by a permutation matrix. For any $\sigma\in\mathfrak{S}_k$, consider the permutation matrix $\mathbf{P}_\sigma=(\delta_{i,\sigma(j)})_{1\leq i,j\leq m}$. We have $\mathbf{P}_\sigma^T=\mathbf{P}_{\sigma^{-1}}$ so that

$$(\widehat{\mathbf{W}}_k\mathbf{P}_\sigma)(\mathbf{P}_\sigma^T\widehat{\mathbf{Z}}_k)=\widehat{\mathbf{W}}_k(\mathbf{P}_\sigma\mathbf{P}_{\sigma^{-1}})\widehat{\mathbf{Z}}_k=\widehat{\mathbf{W}}_k\widehat{\mathbf{Z}}_k.$$

One may refer to (Meilă, 2006; 2007) for a discussion on the permutations of clustering solutions.

Regularization. The constraint of disjoint clusters can be satisfied by forcing the columns of the mixture matrix $\mathbf{W}$ to be orthogonal, *i.e.*, for all $i<j$, $\langle W^i,W^j\rangle=0$. This yields a penalized version of the objective function with a regularization parameter $\lambda>0$

$$\begin{cases}(\widehat{\mathbf{W}}_k,\widehat{\mathbf{Z}}_k)\in\arg\max_{(\mathbf{W},\mathbf{Z})}f_\lambda(\mathbf{W},\mathbf{Z})\\ f_\lambda(\mathbf{W},\mathbf{Z})=Tr(\mathbf{XWZ})/k-\lambda\sum_{i<j}\langle W^i,W^j\rangle\end{cases}$$

with partial derivatives given by

$$\begin{cases}\nabla_{\mathbf{Z}}f_\lambda(\mathbf{W},\mathbf{Z}) &=(\mathbf{XW})^T/k\\ \nabla_{\mathbf{W}}f_\lambda(\mathbf{W},\mathbf{Z}) &=(\mathbf{ZX})^T/k-\lambda\widetilde{\mathbf{W}},\ \widetilde{W}^j=\sum_{i<j}W^i.\end{cases}$$

Update rule. The optimization problem can be addressed using an alternate scheme by computing projected gradient ascent at each iteration

$$\begin{cases}\mathbf{W}_{k+1} &=\Pi_{\mathcal{S}}\left(\mathbf{W}_k+\delta_k^W\nabla_{\mathbf{W}}f_\lambda(\mathbf{W}_k,\mathbf{Z}_k)\right)\\ \mathbf{Z}_{k+1} &=\Pi_{\Delta_m}\left(\mathbf{Z}_k+\delta_k^Z\nabla_{\mathbf{Z}}f_\lambda(\mathbf{W}_{k+1},\mathbf{Z}_k)\right)\end{cases} \tag{8}$$

where $\Pi_{\mathcal{S}}(\cdot),\Pi_{\Delta_m}(\cdot)$ are respectivetly the projection of each column onto a convex set $\mathcal{S}\subset\Delta_p$ and onto the probability simplex Δ_m. The learning rates δ_k^W,δ_k^Z are step sizes found by backtracking line search. The convergence property of the optimization procedure is the same as the convergence of projected gradient descent as detailed in Calamai & Moré (1987); Dunn (1987).

Projection step on $\mathcal{S}$. In order to recover clusters that are not unit sets, we want to avoid the vertices of the simplex. Thus, we perform a projection step $\Pi_{\mathcal{S}}(\cdot)$ of each column of $\mathbf{W}$ onto a convex set $\mathcal{S}$. Several choices are to be considered, as illustrated in Figure 2. Denote $\bar{x}=(1/p,\dots,1/p)$ the barycenter of the probability simplex Δ_p and consider the following manifolds:
(i) ℓ_1 *incircle:* the coordinate permutations of $(0,1/(p-1),\dots,1/(p-1))$ are the centers of the faces of Δ_p and they define a reversed and scaled simplex $\mathcal{S}_p^{\ell_1}$.
(ii) ℓ_2 *incircle:* consider the euclidian ball $B_{2,p}(\bar{x},r)=\{x\in\mathbb{R}^p|\|x-\bar{x}\|_2\leq r\}$. The radius value $r_p=1/\sqrt{p(p-1)}$ yields the ℓ_2 inscribed ball of Δ_p along with $\mathcal{S}_p^{\ell_2}=\Delta_p\cap B_{2,p}(\bar{x},r_p)$.
(iii) $\mathbb{M}$*-set:* The previous manifolds do not scale well as the dimension grows and we shall discuss some theoretical results to see that their hypervolumes become very small. To escape from the curse of dimensionality, we consider the convex set where we cut off the vertices using a threshold τ

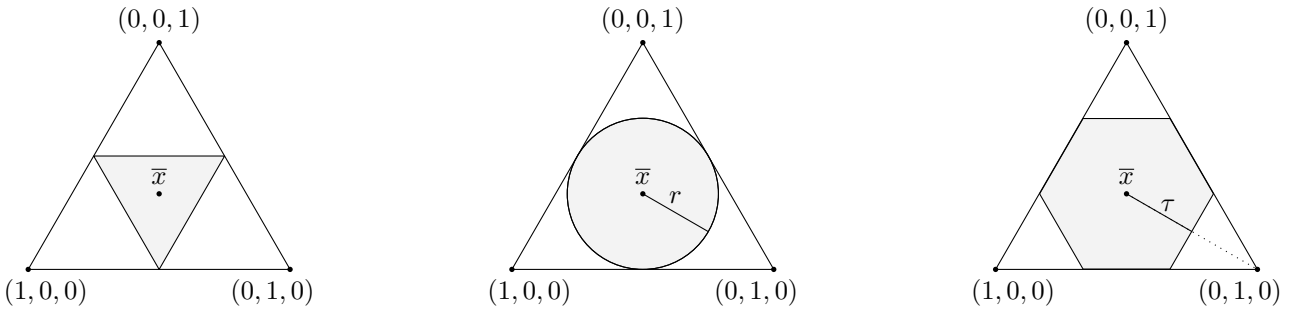

Figure 2. Simplex of $\mathbb{R}^3$ with $\mathcal{S}_3^{\ell_1}$ (left), $\mathcal{S}_3^{\ell_2}$ (center) and the $\mathbb{M}$-set $\mathcal{S}_3^{\tau}$ (right).

of the distance $L = \|\bar{x} - e_j\|_2 = \sqrt{(p-1)/p}$ between the barycenter and a vertex. It is also the intersection of the simplex Δ_p and an ℓ_∞ ball (Warmuth & Kuzmin, 2008; Koolen et al., 2010; Kong et al., 2020). We call this manifold the $\mathbb{M}$-*set* $\mathcal{S}_p^\tau$ defined as

$$\mathcal{S}_p^\tau = \left\{ x \in \Delta_p | \max_{1 \le j \le p} \langle x - \bar{x}, e_j - \bar{x} \rangle \le \tau \|e_j - \bar{x}\|_2 \right\}.$$

Define the radius $r_\infty^p(\tau) = 1 - (1-\tau)(p-1)/p$ then the $\mathbb{M}$-*set* may be seen as the intersection of the simplex with a particular ℓ_∞ ball as

$$\mathcal{S}_p^\tau = \Delta_p \cap B_{\infty,p}(\bar{x}, \tau L) = \Delta_p \cap B_{\infty,p}(0, r_\infty^p(\tau)).$$

The projection onto the simplex is a well-studied subject (Daubechies et al., 2008; Duchi et al., 2008; Chen & Ye, 2011; Condat, 2016). For the projection onto the intersection of convex sets, one can perform a naive approach of alternate projections (Gubin et al., 1967) or some refinements using the idea of Dykstra's algorithm (Dykstra, 1983; Boyle & Dykstra, 1986; Bregman et al., 2003).

3.3. MEXICO Algorithm

Starting from random matrices $(\mathbf{W}_0, \mathbf{Z}_0) \in \mathcal{A}_p^m \times \mathcal{A}_m^k$, the update rule of Eq. (8) returns a pair of matrices $(\mathbf{W}_{mex}, \mathbf{Z}_{mex})$ that are of great interest to analyze the dependence structure of the most extreme data and thus the support of extremes. On the one hand, the mixture matrix $\mathbf{W}_{mex}$ gives insights about the different clusters of features that are large simultaneously. On the other hand, the matrix $\mathbf{Z}_{mex}$ gives information about the probability of belonging to each cluster. Each column $\mathbf{W}_{mex}^j$ represents a cluster K_j and for each sample $\mathbf{X}_i, i \in [\![1, k]\!]$, the j^{th}-row of the column $\mathbf{Z}_{mex}^i$ is the confidence for X_i to belong to the cluster K_j.

A detailed pseudo-code of MEXICO is provided below in Algorithm 1. Since the margins of the data may be unknown, one could work with $\widehat{T}$ which is the empirical counterpart of the Pareto standardization T as detailed in Appendix B.1. The output of the algorithm may be used for feature clustering (FC) or anomaly detection (AD) tasks.

Algorithm 1 MEXICO algorithm

Require: Training data $(X_1, \dots, X_n), 0 < m < p, \lambda > 0$ and rank $k (= \lfloor n\gamma \rfloor)$.
1: Initialize $(\mathbf{W}_0, \mathbf{Z}_0) \in \mathcal{A}_p^m \times \mathcal{A}_m^n$.
2: Standardize the data $\widehat{V}_{(i)} = \widehat{T}(X_{(i)})$ (see Remark 1).
3: Sort training data by decreasing order of magnitude $\|\widehat{V}_{(1)}\|_\infty \ge \dots \ge \|\widehat{V}_{(n)}\|_\infty$.
4: Consider the set of k extreme training data $\widehat{V}_{(1)}, \dots, \widehat{V}_{(k)}$.
5: Compute $(\mathbf{W}_{mex}, \mathbf{Z}_{mex}) \in \arg\max_{(\mathbf{W},\mathbf{Z})} f_\lambda(\mathbf{W}, \mathbf{Z})$ using update rule (8).
6: Given a new input X_{new} standardized as $\widehat{V}_{\text{new}}$ with $\|\widehat{V}_{\text{new}}\|_\infty \ge \|\widehat{V}_{(k)}\|_\infty$, compute $\widetilde{V}_{\text{new}} = \widehat{V}_{\text{new}} \mathbf{W}_{mex}$.
7: Compute predicted cluster $\varphi_0 = \arg\max_{1 \le j \le m} \widetilde{V}_{\text{new}}^j$.
8: **(FC)** Return cluster φ_0.
(AD) Return score $\ell(\widetilde{V}_{\text{new}}, \mathbf{W}_{mex})$.

4. Theoretical Study

This section provides some theoretical results. First, a theoretical analysis of the $\mathbb{M}$-*set* is established. In order to compare the different manifolds of the previous section, we analyze the volume reduction performed in each case. Second, a non-asymptotic bound for the excess risk is detailed.

Theorem 1 (Volume and ratio) *Consider the probability simplex Δ_p and the different manifolds $\mathcal{S}_p^{\ell_1}, \mathcal{S}_p^{\ell_2}, \mathcal{S}_p^\tau$. For any bounded set $\mathcal{D} \subset \mathbb{R}^p$, define its hypervolume $\mathcal{V}ol(\mathcal{D})$ and its ratio $\rho(\mathcal{D})$ as*

$$\mathcal{V}ol(\mathcal{D}) = \int_{\mathbb{R}^p} \mathbb{1}_{\mathcal{D}}(x) \mathrm{d}x, \quad \rho(\mathcal{D}) = \mathcal{V}ol(\mathcal{D}) / \mathcal{V}ol(\Delta_p).$$

$\mathcal{S}$	*Hypervolume* $\mathcal{V}ol(\mathcal{S})$
Δ_p	$\frac{\sqrt{p}}{\Gamma(p)}$
$\mathcal{S}_p^{\ell_1}$	$\frac{\sqrt{p}}{\Gamma(p)}\left(\frac{1}{(p-1)^{(p-1)}}\right)$
$\mathcal{S}_p^{\ell_2}$	$\frac{\sqrt{p}}{\Gamma(p)}\left(\frac{\Gamma(p)}{\Gamma\left(\frac{p+1}{2}\right)}\frac{\pi^{(p-1)/2}}{\sqrt{p^p(p-1)^{(p-1)}}}\right)$
$\mathcal{S}_p^{\tau}$	$\frac{\sqrt{p}}{\Gamma(p)}\left(1-p(1-\tau)^{(p-1)}\left(\frac{p-1}{p}\right)^{(p-1)}\right)$

The corresponding ratios are given by

$$\begin{cases} \rho(\mathcal{S}_p^{\ell_1}) &= \frac{1}{(p-1)^{(p-1)}} \\ \rho(\mathcal{S}_p^{\ell_2}) &= \frac{\Gamma(p)}{\Gamma\left(\frac{p+1}{2}\right)}\frac{\pi^{(p-1)/2}}{\sqrt{p^p(p-1)^{(p-1)}}} \\ \rho(\mathcal{S}_p^{\tau}) &= 1-p\left[(1-\tau)\left(\frac{p-1}{p}\right)\right]^{(p-1)} \end{cases}$$

Moreover, when the dimension grows $p \to +\infty$ and for a fixed $\tau \in (0,1)$, we have $\rho(\mathcal{S}_p^{\ell_1}) \to 0$, $\rho(\mathcal{S}_p^{\ell_2}) \to 0$ and $\rho(\mathcal{S}_p^{\tau}) \to 1$. Among studied subsets, in a high-dimensional setting the $\mathbb{M}$-set is the only one not collapsing towards a unit set.

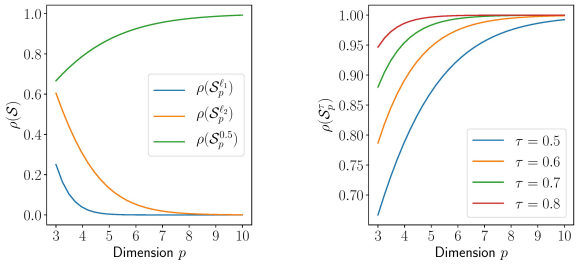

Figure 3. Evolution of the Ratio Volume with dimension p.

Remark 4 *(Selection of τ) With a high reduction of the probability simplex (i.e. $\tau \to 0$), the vertices are avoided but discriminating the clusters is harder as the $\mathbb{M}$-set tends to the barycenter of the simplex. This trade-off motivates the choice of the threshold τ and Figure 3 shows the evolution of the ratios ρ for the different manifolds.*

The following proposition, whose proof is deferred to the supplementary material, shows that MEXICO algorithm can be seen as a contraction mapping.

Proposition 2 (Lipschitz mapping) *Given $x_1, x_2 \in \mathbb{R}_+^p$ with norms greater than $t > 0$ the application $x \mapsto x\mathbf{W}_t^\star$ is $\frac{m}{2}$-lipschitz continuous i.e.,*

$$||x_1\mathbf{W}_t^\star - x_2\mathbf{W}_t^\star||_2 \leq \frac{m}{2}||x_1 - x_2||_2.$$

Moreover, if x_1 and x_2 belong to the same feature cluster K then $||x_1\mathbf{W}_t^\star - x_2\mathbf{W}_t^\star||_2 \leq \frac{1}{2}||x_1 - x_2||_2$. Hence, the transformation induced by MEXICO can be considered as a cluster contraction mapping (Boyd & Wong, 1969).

Finally, the convergence rate of our method can be analyzed through the non-asymptotic bound on the risks (1) and (2). The following Theorem provides an upper bound for the excess risk. The proof is given in the supplementary material.

Theorem 2 (Non-asymptotic bound) *Consider the risk $\mathcal{R}_{t_\gamma}$ defined in (1) associated to the loss function (4) computed on normalized data. Recall that $k = \lfloor n\gamma \rfloor$ and denote by $\mathbf{W}_{mex}$ the mixture matrix obtained by MEXICO. Then for $\delta \in (0,1)$, $n \geq 1$ and $\tau \leq 1$ we have with probability at least $1-\delta$,*

$$\mathcal{R}_{t_\gamma}(\mathbf{W}_{mex}) - \mathcal{R}_{t_\gamma}(\mathbf{W}_{t_\gamma}^\star) \leq \frac{1}{\sqrt{k}} 8\sqrt{2(1-\gamma)\log(4/\delta)} + \frac{1}{k}\left(\frac{16}{3}\log(4/\delta) + 8\sqrt{2(1-\gamma)\log(4/\delta)} + 2\right) + 2r_\infty^p(\tau).$$

The upper bound stated above shows that the convergence rate is of order $O_{\mathbb{P}}(1/\sqrt{k})$ where k is the actual size of the dataset required to estimate the support of extreme. This convergence rate matches the one of Goix et al. (2016).

5. Numerical Experiments

We focus on popular machine learning tasks of *feature clustering* and *anomaly detection* to compare the performance of our algorithm against state-of-the-art methods for extreme events. Since the margins distributions of real-world data are unknown, the rank transformation as described in Remark 1 is considered. For ease of reproducibility, the code is available in the supplementary material.

5.1. Feature Clustering

Consider the feature clustering task where a new extreme sample $X_{\text{new}} \in \mathbb{R}_+^p$ is to be analyzed. Since X_{new} is extreme, our goal is to predict the features that are large simultaneously based on the dependence structure clusters, *i.e.* the clusters given by MEXICO. For that matter, one can compute the transformed sample $\widetilde{X}_{\text{new}} = X_{\text{new}}\mathbf{W}_{mex}$ and assign the predicted cluster of features by $\text{Pred}(X_{\text{new}}) = \arg\max_{1\leq j\leq m} \widetilde{X}_{\text{new}}^j$.

Since MEXICO is an inductive clustering method, we focus on similar clustering algorithms namely spectral clustering (Ding et al., 2005) and spherical K-means (Janßen et al., 2020). Janßen et al. (2020) studied spherical K-means algorithm as a solution to perform clustering in extremes. We consider simulated data from an (asymmetric) logistic distribution where the dependence structure of extremes can be specified (see Appendix B.2). Given the ground truth class samples, we leverage metrics using conditional entropy analysis: Rosenberg & Hirschberg (2007) define the following desirable objectives for any cluster assignment:

Homogeneity (H), each cluster contains only members of a single class; Completeness (C), all members of a given class are assigned to the same cluster; v-Measure (v-M): the harmonic mean of Homogeneity and Completeness.

The parameter setting is the following: dimension $p \in \{75, 100, 150, 200\}$, number of train samples $n_{\text{train}} = 1000$ and test samples $n_{\text{test}} = 100$. We use the metrics implemented by *Scikit-Learn* (Pedregosa et al., 2011). The results, obtained over 100 independently simulated dataset for each value of p, are gathered in Table 1, where the values associated to MEXICO transcribe the best performance between projection method with Dykstra's algorithm and alternating projection. Both methods are detailed in the supplementary material. For each dimension p, bold characters indicate the best method when results are statistically significant using Mann-Whitney and Neyman-Pearson tests.

p	Spectal Clustering (Ding et al., 2005)		
	H	C	v-M
75	0.925± 0.054	0.937±0.040	0.931 ±0.046
100	0.918±0.058	0.934±0.039	0.926±0.048
150	0.889± 0.060	0.925±0.031	0.906± 0.045
200	0.886±0.047	0.928±0.024	0.906±0.034
p	Spherical-Kmeans (Janßen et al., 2020)		
	H	C	v-M
75	0.950±0.034	0.972±0.024	0.961±0.027
100	0.943±0.031	0.967±0.024	0.955±0.026
150	0.940± 0.026	0.962±0.020	0.951±0.022
200	0.940±0.018	0.962±0.014	0.951±0.015
p	MEXICO		
	H	C	v-M
75	**0.978**±0.025	0.976±0.024	**0.977**±0.024
100	**0.978**±0.020	**0.979**±0.021	**0.978**±0.020
150	**0.976**±0.015	**0.980**±0.013	**0.978**±0.014
200	**0.970**±0.015	**0.975**±0.012	**0.972**±0.013

Table 1. Comparison of Homogeneity (H), Completeness (C) and v-Measure (v-M) on Simulated Data.

5.2. Anomaly Detection

To predict whether a new extreme sample $X_{\text{new}} \in \mathbb{R}_+^p$ is an anomaly, one may use the value of the loss function $\ell(X_{\text{new}}, \mathbf{W}_{\text{mex}})$ as an anomaly score. If it is small then the dependence structure of X_{new} is well captured by the mixture $\mathbf{W}_{mex}$ and the behavior is rather *normal*. Conversely, a high value means that X_{new} cannot be approximated by a mixture of $\mathbf{W}_{mex}$ *i.e.* it is more likely to be an outlier. The behavior of the extreme sample X_{new} can be predicted using any decreasing function of the loss function ℓ. In the experiment we use the inverse of the loss though one could consider the opposite of the loss as in (Goix et al., 2016).

We perform a comparison of three algorithms for anomaly detection in extreme regions: Isolation Forest (Liu et al., 2008), DAMEX (Goix et al., 2017) and our method MEXICO. The algorithms are trained and tested on the same datasets, the test set being restricted to extreme regions. Five reference AD datasets are studied: shuttle, forestcover, http, SF and SA. Table 5 in the Appendix provides further dataset details. The experiments are performed in a semi-supervised framework where the training set consists of normal data only. More details about the preprocessing, model tuning and additional results are available in the supplementary material. The results of means and standard deviations are obtained over 100 runs and summarized in Table 2. Better performance are obtained with our anomaly detection approach compared to competing anomaly detection methods.

Dataset	ROC-AUC	AP
	iForest (Liu et al., 2008)	
SA	0.886±0.032	0.879±0.031
SF	0.381±0.086	0.393±0.081
http	0.656±0.094	0.658±0.099
shuttle	0.970±0.020	0.826±0.055
forestcover	0.654±0.096	0.894±0.037
	DAMEX (Goix et al., 2016)	
SA	0.982±0.002	0.938±0.012
SF	0.710±0.031	0.650±0.034
http	0.996±0.002	0.968±0.009
shuttle	0.990±0.003	0.864±0.026
forestcover	0.762±0.008	0.893±0.010
	MEXICO	
SA	0.983±0.031	**0.950**±0.011
SF	**0.892**±0.013	**0.812**±0.016
http	0.997±0.002	**0.972**±0.012
shuttle	0.990±0.003	0.864±0.037
forestcover	**0.863**±0.015	**0.958**±0.006

Table 2. Comparison of Area Under Curve of Receiver Operating Characteristic (ROC-AUC) and Average Precision (AP).

6. Conclusion

Understanding the impact of shocks, *i.e.*, extremely large input values on systems is of critical importance in diverse fields ranging from security or finance to environmental sciences and epidemiology. In this paper, we have developed a a rigorous methodological framework for clustering features in extreme regions, relying on the non-parametric theory of regularly varying random vectors. We illustrated our algorithm performance for both feature clustering and anomaly detection on simulated and real data. Our approach does not scan all the multiple possible subsets and outperforms existing algorithms. Future work will focus on the statistical properties of the developed algorithm by further exploring links with kernel methods.

7. Acknowledgment

Hamid Jalalzai benefitted from the support of the research chair Data Science & Artificial Intelligence for Digitalized Industry and Services at Télécom Paris and by the project HYPATIA, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement n. 835294.

References

Basrak, B., Davis, R. A., and Mikosch, T. A characterization of multivariate regular variation. *Annals of Applied Probability*, pp. 908–920, 2002.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.

Bourbaki, N. *Topologie générale: Chapitres 1 à 4*. Springer Science & Business Media, 2007.

Boyd, D. W. and Wong, J. S. On nonlinear contractions. *Proceedings of the American Mathematical Society*, 20 (2):458–464, 1969.

Boyle, J. P. and Dykstra, R. L. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in order restricted statistical inference*, pp. 28–47. Springer, 1986.

Bregman, L. M., Censor, Y., Reich, S., and Zepkowitz-Malachi, Y. Finding the projection of a point onto the intersection of convex sets via projections onto half-spaces. *Journal of Approximation Theory*, 124(2):194–218, 2003.

Calamai, P. H. and Moré, J. J. Projected gradient methods for linearly constrained problems. *Mathematical programming*, 39(1):93–116, 1987.

Candès, E. J., Romberg, J., and Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Candes, E. J., Romberg, J. K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

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

Chen, Y. and Ye, X. Projection onto a simplex. *arXiv preprint arXiv:1101.6081*, 2011.

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

Chiapino, M., Sabourin, A., and Segers, J. Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22(2):193–222, 2019.

Clémençon, S., Jalalzai, H., Sabourin, A., and Segers, J. Concentration bounds for the empirical angular measure with statistical learning applications. *arXiv preprint arXiv:2104.03966*, 2021.

Clifton, D. A., Hugueny, S., and Tarassenko, L. Novelty detection with multivariate extreme value statistics. *J Signal Process Syst.*, 65:371–389, 2011.

Condat, L. Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 158(1):575–585, Jul 2016. ISSN 1436-4646. doi: 10.1007/s10107-015-0946-6. URL https://doi.org/10.1007/s10107-015-0946-6.

Cooley, D. and Thibaud, E. Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604, 2019.

Cutler, A. and Breiman, L. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

Daubechies, I., Fornasier, M., and Loris, I. Accelerated projected gradient method for linear inverse problems with sparsity constraints. *journal of fourier analysis and applications*, 14(5-6):764–792, 2008.

De Haan, L. and Ferreira, A. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

Devijver, E. et al. Finite mixture regression: a sparse variable selection by model selection for clustering. *Electronic journal of statistics*, 9(2):2642–2674, 2015.

Ding, C., He, X., and Simon, H. D. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining*, pp. 606–610. SIAM, 2005.

Drees, H. and Sabourin, A. Principal component analysis for multivariate extremes. *arXiv preprint arXiv:1906.11043*, 2019.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pp. 272–279. ACM, 2008.

Dunn, J. C. On the convergence of projected gradient processes to singular critical points. *Journal of Optimization Theory and Applications*, 55(2):203–216, 1987.

Dykstra, R. L. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

Einmahl, J. H., Piterbarg, V. I., and De Haan, L. Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics*, 29(5):1401–1423, 2001.

Embrechts, P., Resnick, S. I., and Samorodnitsky, G. Extreme value theory as a risk management tool. *North American Actuarial Journal*, 3(2):30–41, 1999.

Embrechts, P., Klüppelberg, C., and Mikosch, T. *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

Engelke, S. and Hitz, A. S. Graphical models for extremes. *arXiv preprint arXiv:1812.01734*, 2018.

Engelke, S. and Ivanovs, J. Sparse structures for multivariate extremes. *arXiv preprint arXiv:2004.12182*, 2020.

Engelke, S., De Fondeville, R., and Oesting, M. Extremal behaviour of aggregated data with an application to downscaling. *Biometrika*, 106(1):127–144, 12 2018. ISSN 0006-3444. doi: 10.1093/biomet/asy052. URL https://doi.org/10.1093/biomet/asy052.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. *A Geometric Framework for Unsupervised Anomaly Detection*, pp. 77–101. Springer US, 2002.

Friedman, J., Hastie, T., and Tibshirani, R. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

Goix, N., Sabourin, A., and Clémençon, S. Sparse representation of multivariate extremes with applications to anomaly ranking. In *Artificial Intelligence and Statistics*, pp. 75–83, 2016.

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

Gubin, L., Polyak, B. T., and Raik, E. The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24, 1967.

Jalalzai, H., Clémençon, S., and Sabourin, A. On binary classification in extreme regions. In *Advances in Neural Information Processing Systems*, pp. 3092–3100, 2018.

Jalalzai, H., Colombo, P., Clavel, C., Gaussier, É., Varni, G., Vignon, E., and Sabourin, A. Heavy-tailed representations, text polarity classification & data augmentation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Janßen, A., Wan, P., et al. k-means clustering of extremes. *Electronic Journal of Statistics*, 14(1):1211–1233, 2020.

Jessen, H. A. and Mikosch, T. Regularly varying functions. *Publications de l'Institut Mathematique*, 80(94):171–192, 2006.

Karamata, J. Sur un mode de croissance régulière. théorèmes fondamentaux. *Bulletin de la Société Mathématique de France*, 61:55–62, 1933.

KDDCup. The third international knowledge discovery and data mining tools competition dataset. 1999.

Kolesnikov, A., Zhai, X., and Beyer, L. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1920–1929, 2019.

Kong, W., Krichene, W., Mayoraz, N., Rendle, S., and Zhang, L. Rankmax: An adaptive projection alternative to the softmax function. *Advances in Neural Information Processing Systems*, 33, 2020.

Koolen, W. M., Warmuth, M. K., Kivinen, J., et al. Hedging structured concepts. In *COLT*, pp. 93–105. Citeseer, 2010.

Lee, D. D. and Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562, 2001.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Lippmann, R., Haines, J. W., Fried, D., Korba, J., and Das, K. Analysis and results of the 1999 darpa off-line intrusion detection evaluation. In *RAID*, pp. 162–182. Springer, 2000.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008.

Meilă, M. The uniqueness of a good optimum for k-means. In *Proceedings of the 23rd international conference on Machine learning*, pp. 625–632, 2006.

Meilă, M. Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5):873–895, 2007.

Mendez-Civieta, A., Aguilera-Morillo, M. C., and Lillo, R. E. Adaptive sparse group lasso in quantile regression. *Advances in Data Analysis and Classification*, pp. 1–27, 2020.

Meyer, N. and Wintenberger, O. Sparse regular variation. *arXiv preprint arXiv:1907.00686*, 2019.

Niculae, V., Martins, A. F., Blondel, M., and Cardie, C. Sparsemap: Differentiable sparse structured inference. *arXiv preprint arXiv:1802.04223*, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.

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

Resnick, S. I. Point processes, regular variation and weak convergence. *Advances in Applied Probability*, 18(1): 66–138, 1986.

Roberts, S. Novelty detection using extreme value statistics. *IEE P-VIS IMAGE SIGN*, 146:124–129, Jun 1999.

Rosenberg, A. and Hirschberg, J. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420, 2007.

Sabourin, A. and Naveau, P. Bayesian dirichlet mixture model for multivariate extremes: A re-parametrization. *Comput. Stat. Data Anal.*, 71:542–567, 2014.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

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

Stephenson, A. High-dimensional parametric modelling of multivariate extreme events. *Australian & New Zealand Journal of Statistics*, 51:77–88, 2009.

Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A detailed analysis of the kdd cup 99 data set. In *IEEE CISDA*, volume 5, pp. 53–58, 2009.

Tawn, J. Modelling multivariate extreme value distributions. *Biometrika*, 77:245–253, 1990.

Thomas, A., Clemencon, S., Gramfort, A., and Sabourin, A. Anomaly detection in extreme regions via empirical mv-sets on the sphere. In *AISTATS*, pp. 1011–1019, 2017.

Tipping, M. E. and Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Tsaig, Y. and Donoho, D. L. Extensions of compressed sensing. *Signal processing*, 86(3):549–571, 2006.

Vignotto, E. and Engelke, S. Extreme value theory for open set classification–gpd and gev classifiers. *arXiv preprint arXiv:1808.09902*, 2018.

Wang, H. and Leng, C. A note on adaptive group lasso. *Computational statistics & data analysis*, 52(12):5277–5286, 2008.

Warmuth, M. K. and Kuzmin, D. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9 (Oct):2287–2320, 2008.

Wold, S., Esbensen, K., and Geladi, P. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

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

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1):49–67, 2006.

Şimşekli, U., Liutkus, A., and Cemgil, A. T. Alpha-stable matrix factorization. *IEEE Signal Processing Letters*, 22 (12):2289–2293, 2015.