INFORMATIVE CLUSTERS FOR MULTIVARIATE EXTREMES

A PREPRINT

Hamid Jalalzai
LTCI, Télécom Paris
Institut Polytechnique de Paris, France
hamid.jalalzai@telecom-paris.fr

Rémi Leluc
LTCI, Télécom Paris
Institut Polytechnique de Paris, France
remi.leluc@telecom-paris.fr

August 18, 2020

ABSTRACT

Capturing the dependence structure of multivariate extreme data is a major challenge in many fields involving the management of risks that come from multiple sources, e.g., portfolio monitoring, environmental risk management, insurance and anomaly detection. The present paper develops a novel optimization-based approach called MEXICO, standing for Multivariate Extreme Informative Clustering by Optimization. It aims at exhibiting a sparsity pattern within the dependence structure of extremes. This is achieved by estimating some disjoint clusters of features that tend to be large simultaneously through an optimization method on the probability simplex. This dimension reduction technique can be applied to statistical learning tasks such as feature clustering and anomaly detection. Numerical experiments provide strong empirical evidence of the relevance of our approach.

Keywords Multivariate Extreme · Feature Clustering · Sparse Support

1 Introduction

Clustering is essential for exploratory data mining, data structure analysis and a common technique for statistical data analysis. It is widely used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. Many clustering approaches exist with different intrinsic notions of what a cluster is. In the standard setup, the goal is to group objects into subsets, known as clusters, such that objects within a given cluster are more related to one another than the ones from a different cluster. Clustering is already quite well-known (see [4, 27] and references therein) conversely to Extreme Value Theory (EVT) which is a newer field in the machine learning community that has been used in anomaly detection [14, 28, 45, 51], classification [31, 32, 54] or clustering [10, 12, 13, 33] when dedicated to the most extreme regions of the sample space.

Scaling up multivariate EVT is a major challenge when addressing high-dimensional learning tasks. Most multivariate extreme value models have been designed to handle moderate dimensional problems, e.g., with dimension $p \leq 10$. For larger dimensions, simplifying modeling choices are needed, stipulating for instance that only some predefined subgroups of components may be concomitant extremes, or, on the contrary, that all must be [47, 49]. This curse of dimensionality can be explained, in the context of extreme values analysis, by the relative scarcity of extreme data, the computational complexity of the estimation procedure and, in the parametric case, by the fact that the dimension of the parameter space usually grows with that of the sample space.

Recalling the framework of [10, 12, 13], the goal of this paper is to present a novel optimization-based approach for clustering extremes in a multivariate setup. Given $N \geq 1$ i.i.d copies $X_1, \ldots, X_N$ of a heavy-tailed random variable $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$, we want 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 $j \not\in K$ simultaneously remain small. Up to approximately $2^p$ combinations of extreme features are possible and contributions such as [10, 12, 13, 28, 24] tend to identify a smaller number of simultaneous extreme features. Dimensional reduction methods such as principal components analysis and derivatives [16, 21, 52, 55] can be designed to find a lower dimensional subspace where extremes tend to concentrate. Another way of identifying the clusters of features that may jointly be large is to select combinations of extreme features, in the
spirit of archetypes defined in \cite{17}. Following this path, the idea of the present paper is to decompose the \( \ell_1 \)-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 \cite{12,19,25,40}. A broader scope of contributions related to the work detailed in this paper ranges from compressed sensing \cite{8,9,53} and matrix factorization \cite{36,58} to group sparsity \cite{20,48,57}.

The contributions of this paper are: (i) following contribution laid out by \cite{41}, we study at length different manifolds on the probability simplex, (ii) we present a novel optimization-based approach to perform clustering of extreme features in the multivariate set-up with respected regularity property and (iii) we show how to leverage the obtained clusters to detect outliers within the extreme regions in the context of anomaly detection.

The paper is organized as follows, in Section 2 we begin by introducing the EVT background, multivariate set-up and problem of interest. Then we present in Section 3 our optimization-based approach along with its specific details concerning the projection step onto the probability simplex. Section 4 is dedicated to applications in statistical learning, namely feature clustering and anomaly detection. 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 appendix.

2 Probabilistic Framework

Extreme Value Theory (EVT) develops models for learning the unusual rather than the usual, in order to provide a reasonable assessment of the probability of occurrence of rare events. This section introduces the mathematical framework and classical tool such as standardization for the analysis of multivariate extremes.

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 \cite{43} for an excellent account of heavy-tailed distributions and the theory of regularly varying functions.

The following notations are used throughout the paper: \( A_m^p \) is the set of \( p \times m \) matrices valued in \([0, 1]\) where the sum of elements of any column equals 1. Let \( E = [0, \infty[^p \setminus \{0\} \) and \( \Omega_p = \{ x \in \mathbb{R}_+^p : ||x|| \leq 1 \} \) the sphere associated to the norm \( || \cdot || \) and its complementary set \( \Omega_p^c = \mathbb{R}_+^p \setminus \Omega_p \); for \( V \in \mathbb{R}^p \) and \( K \subset [1, p] \), write \( V^{(K)} = (V^j \mathbb{1}_{j \in K}) \).

**Definition 1** (Regular variation \cite{43}) A positive measurable function \( f \) is regularly varying with index \( \alpha \in \mathbb{R} \), notation \( f \in \mathcal{R}_\alpha \) if \( \lim_{x \to +\infty} f(tx)/f(x) = t^\alpha \) 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 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 \cite{43} Section 3.4] and consider the following definition \cite{44} 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 \( f \in \mathcal{R}_{-\alpha} \) and a nonzero Radon measure \( \mu \) on \( E \) such that

\[
\lim_{t \to +\infty} t^{-\alpha} \mathbb{P}(t^{-1} X \in A) \xrightarrow{t \to +\infty} \mu(A),
\]

where \( A \subset E \) is any Borel set such that \( 0 \not\in \partial A \) and \( \mu(\partial A) = 0 \).

**Standardization.** Let \( F \) be the joint cumulative distribution function (c.d.f) of \( X \) and \( F^j \) be the marginal c.d.f of \( X^j \) with \( j \in [1, p] \). The tails of the marginals may differ and it is convenient, in practice, to work with marginally standardized variables. In other words, we separate the margins from the dependence structure in the description of the joint distribution of \( X \) to compare the different features \( X^j \). For that purpose, we consider the Pareto scaling \( T : \mathbb{R}^p \to \mathbb{R}_+^p \) defined by

\[
\forall x \in \mathbb{R}^p, \forall j \in [1, p], \quad T^j(x^j) = 1/(1 - F^j(x^j)) \in [1, +\infty].
\]
This transformation produces a vector $\tilde{V} \overset{\text{def}}{=} T(X) = (T(X_1), \ldots, T(X_p))$ where each marginal follows a Pareto distribution, i.e., $\mathbb{P}(V^j > t) = t^{-1}$ for all $t > 1$. In such a case, a particular regularly varying function is $f(t) = t^{-1}$ and $\tilde{V} \in \mathbb{R}_+^p$ is regularly varying with tail index $\alpha = 1$ (see [11] Remark 1). The limiting measure $\mu$ is homogeneous and known as the exponent measure.

In practice, the marginal functions $F^j$ are unknown and need to be approximated. Assume that $N \geq 1$ i.i.d copies $X_1, \ldots, X_N$ of a heavy-tailed random variable $X \in \mathbb{R}^p$ are available. When the margins are unknown, the Pareto scaling can be approximated by the rank transformation $\tilde{T} : \mathbb{R}^p \to \mathbb{R}_+^p$ relying on the empirical marginal c.d.f. denoted by $\hat{F}^j(x) = (1/N + 1) \sum_{i=1}^N I\{X_i^j \leq x\}$, with $x \in \mathbb{R}$. The empirical version $\hat{V}$ of $V$ from Equation (1) is defined by

$$\hat{V}_i = \hat{T}(X_i) = \left(\frac{1}{1-\hat{F}^1(X_i^1)}, \ldots, \frac{1}{1-\hat{F}^p(X_i^p)}\right), \quad \forall i \in [1, N]. \quad (2)$$

The extreme region is then selected by choosing all vectors with norm larger than a fixed threshold $t > 0$, i.e., the data extreme vectors $\hat{V}$ such that $\|\hat{V}\| > t$, yielding to $n$ samples considered as extremes. 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 limit measure definition [3].

Note that this rank standardization is commonly used in multivariate EVT to study the dependence structure of extremes (see [3] and references therein) and avoids any further marginal distributions assumptions. The resulting feature variables of $\hat{V}$ are not independent and the remaining goal is to discover the dependence structure of standardized extremes.

**Problem statement.** We consider a vector $V \in \mathbb{R}_+^p$ whose features come from a mixture of extreme values and we would like to find clusters of features that get excited together. We seek $m \geq 2$ clusters $K_1, \ldots, K_m$ with $m < p$ such that all features in a same subset may be large together. Unit sets are not relevant for clustering so we assume that each cluster is of size at least 2. We also want clusters that are disjoint, i.e., for all $i \neq j$, $K_i \cap K_j = \emptyset$. This choice is motivated to reach a representation of interest, e.g., diversity for portfolio in finance or clusters for smart grids in wireless technologies. In the remaining of this paper, $V \in \mathcal{M}_{n,m}([1, +\infty])$ corresponds to all the samples $X_1, \ldots, X_N$ after the rank standardization and selection of extremes. With this notation, we have $n$ samples $V_1, \ldots, V_n \in \mathbb{R}_+^p$ that are i.i.d. copies of the vector $V$ and for all $i \in [1, n]$, we search a subset $K$ of features such that the $\ell_1$-norms of $V_i$ and its restriction $V_i^K$ are almost equal $\|\hat{V}_i\|_1 \approx \|V_i^K\|_1$.

### 3 Optimization Problem

**Features mixtures.** In order to recover the clusters, we consider mixtures of the components of each sample. The true number of clusters is unknown so we can only have a guess and search for $m$ clusters. We consider the probability simplex defined on the positive orthant $\mathbb{R}_+^p$ by

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

and let $W \in A_m^n$ with $m < p$ be a mixture matrix. We denote by $\tilde{V} = VW \in \mathcal{M}_{n,m}(\mathbb{R}_+)$ the transformed matrix. The following theorem ensures the preservation of the regularly varying behavior and points out the behavior of the limiting measures.

**Theorem 1** Let $V = T(X) \in \mathbb{R}_+^p$ coming from a Pareto scaling and $W \in A_m^n$ a mixture matrix with $1 < m \leq p$. Then the transformed vector $\tilde{V} = VW \in \mathbb{R}_+^m$ is regularly varying with tail index $\alpha = 1$. Denote $\mu$ (resp. $\tilde{\mu}$) the limiting measure of $V$ (resp. $\tilde{V}$), then we have $\tilde{\mu}(\Omega_m^\alpha) \leq \mu(\Omega_p^\alpha)$.

**Remark 1** (Selection of $m$) In view of Theorem 1, in practice, the required dimension $m < p$ can be seen as the smallest value $m$ such that the empirical version of $\tilde{\mu}(\Omega_m^\alpha)$ is arbitrarily close to the empirical version of $\mu(\Omega_p^\alpha)$. In that way, the $m$ selected clusters provide a good enough representation of the dependency between features.

**Loss function.** Each column $W^j$ for $j \in [1, m]$ is modelling a mixture of components and represents a cluster $K_j$. For any sample $V_i$, $i \in [1, n]$, we want to find a mixture that gives a good approximation in $\ell_1$-norm, i.e., we seek a column $j \in [1, m]$ for which $\tilde{V}_i^j$ is the closest to $||V_i||_1$. In other words, we need to find $j \in [1, m]$ in order to minimize the score function $\gamma$ defined as

$$\forall (i, j) \in [1, n] \times [1, m], \quad \gamma(W, V_i, j) = ||V_i||_1 - \tilde{V}_i^j.$$
For each sample \( V_i \), we need to minimize the loss function defined by
\[
\mathcal{L}(W, V_i) = \min_{1 \leq j \leq m} \gamma(V_i, W, j) = \min_{1 \leq j \leq m} \left( ||V_i||_1 - \tilde{V}_i^j \right).
\]
The optimization problem consists in finding a mixture matrix \( W^* \) minimizing the global loss
\[
W^* \in \arg \min_{W \in \mathcal{A}_n^m} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(W, V_i).
\]
Note that \( \mathcal{A}_n^m \) is a closed and bounded set hence compact \(\overline{\mathcal{A}}\) thus there exists at least one solution which can be reached. Equation (3) is composed of two minimization problems and can be rewritten as
\[
W^* \in \arg \max_{W \in \mathcal{A}_n^m} \frac{1}{n} \sum_{i=1}^{n} \max_{1 \leq j \leq m} \tilde{V}_i^j.
\]
The index of the column representing a good mixture can be defined with the mapping
\[
\varphi : [1, n] \rightarrow [1, m], \quad \varphi(i) = \arg \max_{1 \leq j \leq m} \tilde{V}_i^j
\]
and the optimization problem becomes
\[
W^* \in \arg \max_{W \in \mathcal{A}_n^m} \frac{1}{n} \sum_{i=1}^{n} (VW)^\varphi(i) = \arg \max_{W \in \mathcal{A}_n^m} \frac{1}{n} \sum_{i=1}^{n} e_i(VW)e^{\varphi(i)}.
\]
**Simple example.** To have a better understanding of the optimization problem, we consider a simple example to show how the matrix \( W \) is recovering the different clusters. Assume that the vector \( V \in \mathbb{R}_+^m \) is exactly coming from a mixture of \( m \) disjoint clusters \( K_1, \ldots, K_m \), and for each sample \( V_i \), there exists \( K_j \) such that \( ||V_i^{(K_j)}||_1 = ||V_i||_1 \). 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 \( W^* \) whose columns are the vectors \( U^j \). Indeed, the transformed data matrix is \( \tilde{V} = VW \) and for any sample \( V_i \) that comes from a cluster \( K_j \), we have
\[
\forall l \neq j, \quad \tilde{V}_i^j = V_i U^j = V_i^{(K_j)} U^j \geq V_i U^l = \tilde{V}_i^l.
\]
Taking \( \varphi(i) = \arg \max_{1 \leq l \leq m} \tilde{V}_i^l \) exactly recovers the cluster of index \( j = \varphi(i) \). In the case where the large features of the different sample \( V_i \) are all equal, then the columns of the mixture matrix \( W \) tend exactly to uniform vectors with restricted support. Now, if one of the excited feature is slightly bigger than the other then the associated column of \( W \) tends to a vertex of the simplex.

**Problem relaxation.** One can directly solve the linear program (4) but this formulation suffers from drawbacks. First, the solution provided tends to be very sparse since it would belong to a vertex of the simplex. Then it involves the search of the mapping \( \varphi \) among all the possible combinations which can be prohibited when \( n \) or \( p \) increases. Thus, one can solve a relaxed version of (4) by introducing another matrix of mixtures \( Z \in \mathcal{A}_n^m \). The relaxed problem is
\[
(W^*, Z^*) \in \arg \max_{(W,Z) \in \mathcal{A}_n^m \times \mathcal{A}_n^m} \frac{1}{n} \sum_{i=1}^{n} V_i W Z^i.
\]
**Optimization problem.** We recognize the trace operator which is linear and can define an objective function \( f : \mathcal{A}_n^m \times \mathcal{A}_n^m \rightarrow \mathbb{R} \) that we need to maximize:
\[
\begin{cases}
(W^*, Z^*) \in \arg \max_{(W,Z) \in \mathcal{A}_n^m \times \mathcal{A}_n^m} f(W, Z) \\
f(W, Z) = Tr(VWZ)/n
\end{cases}
\]
The objective function \( f \) is bilinear in finite dimension hence continuous. Since maximization occurs on compact sets, there is at least one solution \( (W^*, Z^*) \). However, it is not unique since any column-permutation of \( W^* \) along with the associated row-permutation of \( Z^* \) is also a valid solution.

**Regularization.** The constraint of disjoint clusters can be satisfied by forcing the columns of the mixture matrix \( W \) 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 \( \lambda > 0 \)
\[
\begin{cases}
(W^*, Z^*) \in \arg \max_{(W,Z) \in \mathcal{A}_n^m \times \mathcal{A}_n^m} f_\lambda(W, Z) \\
f_\lambda(W, Z) = Tr(VWZ)/n - \lambda \sum_{i<j} \langle W^i, W^j \rangle
\end{cases}
\]
with partial derivatives given by
\[
\begin{align*}
\nabla_Z f_\lambda(W, Z) &= (ZW)^T/n \\
\nabla_W f_\lambda(W, Z) &= (ZW)^T/n - \lambda \tilde{W}, \quad \tilde{W} = \sum_{i<j} W^i.
\end{align*}
\]

**Update rule.** The optimization problem can be addressed using an alternate scheme by computing projected gradient ascent at each iteration
\[
\begin{align*}
W_{k+1} &= \Pi_S (W_k + \delta^W_k \nabla W f_\lambda(W_k, Z_k)) \\
Z_{k+1} &= \Pi_{\Delta_m} (Z_k + \delta^Z_k \nabla Z f_\lambda(W_{k+1}, Z_k))
\end{align*}
\]

where \(\Pi_S(\cdot)\), \(\Pi_{\Delta_m}(\cdot)\) are respectively the projection of each column onto a convex set \(S \subset \Delta_p\) and onto the probability simplex \(\Delta_m\). The learning rates \(\delta^W_k, \delta^Z_k\) are step sizes found by backtracking line search.

**Projection step on \(S\).** 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 \(\Pi_S(\cdot)\) of each column of \(W\) onto a convex set \(S\). Several choices are to be considered, as illustrated in Figure 1. Denote \(\tilde{x} = (1/p, \ldots, 1/p)\) the barycenter of the probability simplex \(\Delta_p\) and consider the following manifolds:

(i) **\(\ell_1\ incircle**; the coordinate permutations of \((0, 1/(p-1), \ldots, 1/(p-1))\) are the centers of the faces of \(\Delta_p\) and they define a reversed and scaled simplex \(S^\ell_1\).

(ii) **\(\ell_2\ incircle**; consider the euclidian ball \(B_{2,p}(\tilde{x}, r) = \{x \in \mathbb{R}^p \|x - \tilde{x}\|_2 \leq r\}\). The radius value \(r_p = 1/\sqrt{p(p-1)}\) yields the \(\ell_2\) inscribed ball of \(\Delta_p\) along with \(S^\ell_2 = \Delta_p \cap B_{2,p}(\tilde{x}, r_p)\).

(iii) **Mexican 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 \(\tau\) of the distance \(L = \|x - e_j\|_2 = \sqrt{(p-1)/p}\) between the barycenter and a vertex. It is also the intersection of the simplex \(\Delta_p\) and an \(\ell_\infty\) ball. We call this manifold the Mexican set \(S^\tau\) defined as
\[
S^\tau = \left\{ x \in \Delta_p \left| \max_{1 \leq j \leq p} \left( x - \tilde{x}, \frac{e_j - \tilde{x}}{\|e_j - \tilde{x}\|_2} \right) \leq \tau \right. \right\}.
\]

Define \(r^\tau_p(\tau) = 1 - (1 - \tau)(p-1)/p\) then we also have the relation
\[
S^\tau = \Delta_p \cap B_{\ell_\infty,p}(\tilde{x}, \tau L) = \Delta_p \cap B_{\ell_\infty,p}(0, r^\tau_p(\tau)).
\]

The projection onto the simplex is a well-studied subject \([11, 15, 18, 22]\). For the projection onto the intersection of convex sets, one can perform a naive approach of alternate projections \([40]\) or some refinements using the idea of Dykstra’s algorithm \([6, 7, 23]\).

![Figure 1: Simplex of \(\mathbb{R}^3\) with \(S^\ell_1\) (left), \(S^\ell_2\) (center) and the Mexican set \(S^\tau\) (right).](image)

**Theorem 2** (Volumes and ratios) Consider the probability simplex \(\Delta_p\) and the different manifolds \(S^\ell_1, S^\ell_2, S^\tau\). Denote \(\Gamma\) the Euler function, with \(\Gamma(p) = (p-1)!\). For any bounded set \(D \subset \mathbb{R}^p\), define its hypervolume \(Vol(D)\) and its ratio \(\rho(D) = Vol(D)/Vol(\Delta_p)\).

Moreover, when the dimension grows \(p \to +\infty\) and for a fixed \(\tau \in (0, 1)\), we have \(\rho(S^\ell_1) \to 0\), \(\rho(S^\ell_2) \to 0\), and \(\rho(S^\tau) \to 1\).

**Remark 2** (Selection of \(\tau\)) With a high reduction of the probability simplex, the vertices are avoided but the clusters are more difficult to discriminate since the Mexican set tends to the barycenter of the simplex. This trade-off motivates the choice of the threshold \(\tau\) and Figure 2 shows the evolution of the ratios \(\rho\) for the different manifolds.
### 4 Statistical Learning Applications

Starting from random matrices \((W_0, Z_0) \in A^m_p \times A^n_p\), the algorithm of Equation (6) returns a pair of matrices \((W\)\text{\textsubscript{mex}}, Z\text{\textsubscript{mex}}) that are of great interest to analyze the dependence structure of the extreme data. On the one hand, the mixture matrix \(W\text{\textsubscript{mex}}\) gives insights about the different clusters of features that are large simultaneously. On the other hand, the matrix \(Z\text{\textsubscript{mex}}\) gives information about the probability of belonging to each cluster. Indeed, those matrices are trained on the data matrix \(V\) so that each column \(W\text{\textsubscript{mex}}\) represents a cluster \(K_j\) and for each sample \(V_i, i \in [1, n]\), the \(j\text{\textsuperscript{th}}\)-row of the column \(Z\text{\textsubscript{mex}}\) is the confidence of belonging to the cluster \(K_j\).

#### Features Clustering.
Consider the features clustering task where we receive a new extreme sample \(V\text{\textsubscript{new}} \in \mathbb{R}^p\) and need to predict the cluster where its large features are drawn. To assess the dependency structure of this new sample, one can compute the transformed sample \(V\text{\textsubscript{new}}\) and assign the predicted cluster by

\[
\tilde{V}\text{\textsubscript{new}} = V\text{\textsubscript{new}} W\text{\textsubscript{mex}}, \quad \text{Pred}(V\text{\textsubscript{new}}) = \arg \max_{1 \leq j \leq m} \tilde{V}^j\text{\textsubscript{new}}.
\]

#### Anomaly Detection.
Consider now the anomaly detection task where we receive a new extreme sample \(V\text{\textsubscript{new}} \in \mathbb{R}^p\) and need to predict whether it is an anomaly or not. One can look at the score function evaluated at the new sample \(\gamma (W\text{\textsubscript{mex}}, V\text{\textsubscript{new}}, \varphi\text{\textsubscript{new}})\) where \(\varphi\text{\textsubscript{new}} = \arg \max_{1 \leq j \leq m} \tilde{V}^j\text{\textsubscript{new}}\). If this score is small then it means that the dependency structure of \(V\text{\textsubscript{new}}\) is well captured by the mixture \(W\text{\textsubscript{mex}}\) and the behavior is rather normal that unusual. Similarly, a high value of this score means that \(V\text{\textsubscript{new}}\) cannot be well explained by any mixture of \(W\text{\textsubscript{mex}}\) and therefore it is more likely to be an outlier. Based on that remark, it is easy to make a prediction for the behavior of the extreme sample \(V\text{\textsubscript{new}}\) using a decreasing function of the score value.

**Remark 3** (On selection of \(k\) in Algorithm 7) Determining \(k\) is a central bias variance trade-off of Extreme Value analysis (See e.g. [28] 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. In practice, a conventional choice is \(k = \sqrt{N}\).
We perform a comparison of three algorithms: Isolation Forest [39], Damex [29] and Mexico. The algorithms are trained and tested on the same datasets, the test set being restricted to extreme regions. Five reference AD datasets are considered: shuttle, forestcover, http, SF and SA. The setting is detailed in Table 2. The experiments are performed in a semi-supervised framework where the training set consists of normal data only. More details about the preprocessing and additional results are available in the appendix.

The results of means and standard deviations, obtained over 100 runs, are gathered in Table 3 and reveal the good performance of our approach.
| Dataset   | Size  | Anomalies | τ | λ |
|-----------|-------|-----------|---|---|
| SF        | 73 237| 3298 (4.5%)| 0.8| 10 |
| SA        | 100 655| 3377 (3.4%)| 0.7| 5  |
| http      | 58 725| 2209 (3.8%)| 0.5| 10 |
| shuttle   | 49 097| 3511 (7.2%)| 0.7| 5  |
| forestcover| 286 048| 2747 (0.9%)| 0.7| 5  |

Table 2: Description of each dataset and hyperparameters of Mexico for anomaly detection.

| Dataset   | iForest [39] | ROC-AUC | AP  |
|-----------|--------------|---------|-----|
| SF        | 0.381±0.086  | 0.393±0.081|     |
| SA        | 0.886±0.032  | 0.879±0.031|     |
| 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|     |

| Dataset   | Damex [28] | ROC-AUC | AP  |
|-----------|------------|---------|-----|
| SF        | 0.710±0.031| 0.650±0.034|     |
| SA        | 0.983±0.031| 0.950±0.011|     |
| http      | 0.997±0.002| 0.972±0.012|     |
| shuttle   | 0.990±0.003| 0.864±0.037|     |
| forestcover| 0.997±0.002| 0.972±0.012|     |

| Dataset   | Mexico | ROC-AUC | AP  |
|-----------|--------|---------|-----|
| SF        | 0.892±0.013| 0.812±0.016|     |
| SA        | 0.983±0.031| 0.950±0.011|     |
| http      | 0.997±0.002| 0.972±0.012|     |
| shuttle   | 0.990±0.003| 0.864±0.037|     |
| forestcover| 0.997±0.002| 0.972±0.012|     |

Table 3: Comparison of Area Under Curve of Receiver Operating Characteristic (ROC-AUC) and Average Precision (AP) from prediction scores of each method on different anomaly detection datasets.

6 Conclusion

Understanding the impact of shocks, *i.e.*, extremely large input values on systems is of critical importance in diverse fields, *e.g.*, security, finance, environmental sciences, epidemiology. In this paper, we have developed a rigorous methodological framework for clustering in extreme regions, relying on the non-parametric theory of regularly varying random vectors, and illustrated its 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.

From a broader perspective, extreme data may have dramatic consequences and any clustering algorithm in such a setting should be used with great caution. Note that the purpose of MEXICO is to provide informative clusters of features although no guarantees on its robustness are provided so far. Clustering rare events is the cornerstone of many applications and may have huge social consequences, *e.g.*, a river dam failure in hydrological sciences or a miscarriage of justice when dealing with homeland security. Finally, recovering clusters of data concerning sick patients at an early stage of a global pandemic is the key to slow down the resulting epidemic. In this way, future work will focus on the statistical properties and guarantees of the developed algorithm by further exploring links with kernel methods.

References

[1] Bojan Basrak, Richard A Davis, and Thomas Mikosch. A characterization of multivariate regular variation. *Annals of Applied Probability*, pages 908–920, 2002.
[2] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.
[3] Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef L Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.
[4] C. M Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.
[5] Nicolas Bourbaki. *Topologie générale: Chapitres I à 4*. Springer Science & Business Media, 2007.
[6] James P Boyle and Richard L Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in order restricted statistical inference*, pages 28–47. Springer, 1986.
[7] Lev M Bregman, Yair Censor, Simeon Reich, and Yael Zepkowitz-Malachi. 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.
[8] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.
[9] Emmanuel J Candes, Justin K Romberg, and Terence Tao. 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.
[10] Emilie Chautru et al. Dimension reduction in multivariate extreme value analysis. *Electronic journal of statistics*, 9(1):383–418, 2015.

[11] Yunmei Chen and Xiaojing Ye. Projection onto a simplex. *arXiv preprint arXiv:1101.6081*, 2011.

[12] Maël Chiapino and Anne Sabourin. 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, 2016.

[13] Maël Chiapino, Anne Sabourin, and Johan Segers. Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22(2):193–222, 2019.

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

[15] Laurent Condat. Fast projection onto the simplex and the $\ell_1$ ball. *Mathematical Programming*, 158(1):575–585, Jul 2016.

[16] Daniel Cooley and Emeric Thibaud. Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604, 2019.

[17] Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

[18] Ingrid Daubechies, Massimo Fornasier, and Ignace Loris. Accelerated projected gradient method for linear inverse problems with sparsity constraints. *Journal of Fourier Analysis and Applications*, 14(5-6):764–792, 2008.

[19] Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

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

[21] Holger Drees and Anne Sabourin. Principal component analysis for multivariate extremes. *arXiv preprint arXiv:1906.11043*, 2019.

[22] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the $l_1$-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.

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

[24] Sebastian Engelke and Adrien S Hitz. Graphical models for extremes. *arXiv preprint arXiv:1812.01734*, 2018.

[25] Sebastian Engelke and Jevgenijs Ivanovs. Sparse structures for multivariate extremes. *arXiv preprint arXiv:2004.12182*, 2020.

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

[27] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001.

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

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

[30] LG Gubin, Boris T Polyak, and EV Raik. The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24, 1967.

[31] Hamid Jalalzai, Stephan Clémençon, and Anne Sabourin. On binary classification in extreme regions. In *Advances in Neural Information Processing Systems*, pages 3092–3100, 2018.

[32] Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. Heavy-tailed representations, text polarity classification & data augmentation. *arXiv preprint arXiv:2003.11593*, 2020.

[33] Anja Janßen, Phyllis Wan, et al. $k$-means clustering of extremes. *Electronic Journal of Statistics*, 14(1):1211–1233, 2020.

[34] Jovan Karamata. *Sur un mode de croissance régulière, théorèmes fondamentaux*. *Bulletin de la Société Mathématique de France*, 61:55–62, 1933.
[35] KDDCup. The third international knowledge discovery and data mining tools competition dataset. 1999.

[36] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In Advances in neural information processing systems, pages 556–562, 2001.

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

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

[39] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining, pages 413–422. IEEE, 2008.

[40] Nicolas Meyer and Olivier Wintenberger. Sparse regular variation. arXiv preprint arXiv:1907.00686, 2019.

[41] Vlad Niculae, André FT Martins, Mathieu Blondel, and Claire Cardie. Sparsemap: Differentiable sparse structured inference. arXiv preprint arXiv:1802.04223, 2018.

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

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

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

[45] S.J. Roberts. Novelty detection using extreme value statistics. IEEE P-Vis IMAGE SIGN, 146:124–129, Jun 1999.

[46] Andrew Rosenberg and Julia Hirschberg. 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), pages 410–420, 2007.

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

[48] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. Journal of computational and graphical statistics, 22(2):231–245, 2013.

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

[50] M. Tavallaeae, E. Bagheri, W. Lu, and A.A. Ghorbani. A detailed analysis of the kdd cup 99 data set. In IEEE CISDA, volume 5, pages 53–58, 2009.

[51] Albert Thomas, Stephan Clemencon, Alexandre Gramfort, and Anne Sabourin. Anomaly detection in extreme regions via empirical mv-sets on the sphere. In AISTATS, pages 1011–1019, 2017.

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

[53] Yaakov Tsaig and David L. Donoho. Extensions of compressed sensing. Signal processing, 86(3):549–571, 2006.

[54] Edoardo Vignotto and Sebastian Engelke. Extreme value theory for open set classification–gpd and gev classifiers. arXiv preprint arXiv:1808.09902, 2018.

[55] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3):37–52, 1987.

[56] Kenji Yamanishi, Jun-Ichi Takeuchi, Graham Williams, and Peter Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Mining and Knowledge Discovery, 8(3):275–300, 2004.

[57] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67, 2006.

[58] U. Simşekli, A. Liutkus, and A. T. Cemgil. Alpha-stable matrix factorization. IEEE Signal Processing Letters, 22(12):2289–2293, 2015.
APPENDIX

Section A gathers the proofs of the two theorems and Section B is dedicated to numerical experiments: the preprocessing of the data and additional results. In Section C we present further numerical experiments with a visualization of the clusters found by the algorithm.

A Proofs of Theorems

A.1 Proof of Theorem I

PROOF. Assume that \( V = T(X) \in \mathbb{R}^p_+ \) is coming from a Pareto scaling. Then each marginal \( V^j \) for \( j = 1, \ldots, p \) follows a Pareto distribution and \( V \) is a regularly varying random vector with tail index 1 (see [31]). Using the characterization of [1], we have the following equivalence between the behavior of the vector and its components

\[
(V \text{ is regularly varying}) \iff (\forall u \in \mathbb{R}^p, \langle u, V \rangle \text{ is univariate regularly varying})
\]

Each column \( \tilde{V}^j \) of \( \tilde{V} = VW \) is given by the linear combination \( \tilde{V}^j = \sum_{k=1}^p V^k W^j_k \). Therefore, any linear combination of the form \( \langle \tilde{u}, \tilde{V} \rangle \) with \( \tilde{u} \in \mathbb{R}^m \) is actually a linear combination of the form \( \langle u, V \rangle \). Indeed, we have for \( \tilde{u} \in \mathbb{R}^m \),

\[
\langle \tilde{u}, \tilde{V} \rangle = \sum_{j=1}^m \tilde{u}_j \tilde{V}^j = \sum_{j=1}^m \tilde{u}_j \left( \sum_{k=1}^p V^k W^j_k \right) = \sum_{k=1}^p \left( \sum_{j=1}^m \tilde{u}_j W^j_k \right) V^k.
\]

Because \( V \) is regularly varying then any linear combination of the form \( \langle \tilde{u}, \tilde{V} \rangle \) is univariate regularly varying, which exactly means, using the equivalence, that \( \tilde{V} \) is a regularly varying random vector. To find the tail index of the transformed vector, we rely on the following Lemma.

**Lemma 1** Let \( V \in \mathbb{R}^p_+ \) be a random vector with each component following a Pareto distribution and \( W \in A^m_p \) a mixture matrix. Then each marginal of the transformed vector \( VW \) follows a Pareto distribution.

PROOF. For all \( j \in [1, p] \), the marginal \( V^j \) follows a Pareto distribution. In particular, this is true for the smallest marginal \( V_m = \min_{1 \leq j \leq m} V^j \) and largest marginal \( V_M = \max_{1 \leq j \leq m} V^j \). We have for all \( k \in [1, p] \), \( V_m \leq V^k \leq V_M \). Since the features of the transformed vector \( \tilde{V} = VW \) are given by \( \tilde{V}^j = \sum_{k=1}^p V^k W^j_k \) with \( W^j \in \Delta_p \), we can sum these inequalities and get that \( V_m \leq \tilde{V}^j \leq V_M \). Therefore, we have

\[
\forall t > 1, \quad \mathbb{P}(V_m > t) \leq \mathbb{P}(\tilde{V}^j > t) \leq \mathbb{P}(V_M > t).
\]

Because \( V_m \) and \( V_M \) follow Pareto distribution then so does \( \tilde{V}^j \) with \( \mathbb{P}(\tilde{V}^j > t) = 1/t \).

Since the random vectors \( V \) and \( \tilde{V} \) are regularly varying, we have the existence of nonzero Radon measures \( \mu \) and \( \tilde{\mu} \) that are independent of the considered norm (see [3]). Moreover, in virtue of Lemma 1, the marginals of \( V \) and \( \tilde{V} \) follow Pareto distributions so their tail index is equal to 1 (see [31]). Consider the complementary set of the unit sphere, defined by \( \Omega_m^c = \{ x \in \mathbb{R}^m_+, \| x \| > 1 \} \). We have by definition

\[
\tilde{\mu}(\Omega_m^c) = \lim_{t \to \infty} t \mathbb{P} \left( t^{-1} \tilde{V} \in \Omega_m^c \right) = \lim_{t \to \infty} t \mathbb{P} \left( \| \tilde{V} \| > t \right).
\]

Using that \( \| \tilde{V} \|_\infty = \max_{1 \leq j \leq m} \tilde{V}^j = \max_{1 \leq j \leq m} \left( \sum_{k=1}^p V^k W^j_k \right) \) and \( W^j_k \in [0, 1] \), we have

\[
\mathbb{P} \left( \| \tilde{V} \|_\infty > t \right) = \mathbb{P} \left( \max_{1 \leq j \leq m} \tilde{V}^j > t \right) \leq \mathbb{P} \left( \sum_{k=1}^p V^k > t \right).
\]

We recognize the \( \ell_1 \)-norm of the random vector \( V \in \mathbb{R}^p_+ \) and obtain

\[
\forall t > 1, \quad t \mathbb{P} \left( \| \tilde{V} \|_\infty > t \right) \leq t \mathbb{P} \left( \| V \|_1 > t \right).
\]

Taking the limit \( t \to \infty \) on both sides provides the desired result \( \tilde{\mu}(\Omega_m^c) = \mu(\Omega_p^c) \).

11
A.2 Proof of Theorem 2

PROOF. First, recall the hypervolume of the p-simplex with side length a and the hypervolume of the Euclidian ball of radius R in dimension p,

\[ \text{Vol}(\Delta_p, a) = \frac{\sqrt{p}}{(p-1)!} \left( \frac{a}{\sqrt{2}} \right)^{p-1}, \quad \text{Vol}(B_{2,p}(0, R)) = \frac{\pi^{p/2}R^p}{\Gamma\left(\frac{p}{2} + 1\right)}. \]

**Probability simplex \( \Delta_p \).** The probability simplex we consider has a side length of \( a = \sqrt{2} \) which gives the value of \( \text{Vol}(\Delta_p) \).

\( \ell_1 \)-incircle. Regarding the \( \ell_1 \)-ball, it is the scaled simplex whose side length is given by the distance between two face centers of \( \Delta_p \). This length is equal to \( \sqrt{2}/(p-1) \) and we deduce the volume \( \text{Vol}(S\ell_1^p) \).

\( \ell_2 \)-incircle. For the \( \ell_2 \)-ball, denote \( B = (e_1, \ldots, e_p) \) the canonical basis and let \( x \in S\ell_2^p, x = \sum_{i=1}^p \langle x, e_i \rangle e_i = \sum_{i=1}^p x_i e_i \). The vector \( e'_p = \sqrt{p} \bar{x} = (1/\sqrt{p}, \ldots, 1/\sqrt{p}) \) is unitary and orthogonal to the simplex \( \Delta_p \) with \( \Delta_p \subset \text{Span}(e'_1, \ldots, e'_p) \). We have \( \langle x, e'_p \rangle = 0 \) and we can complete the vector \( e'_p \) into an orthonormal basis \( B' = (e_1, \ldots, e'_p) \) with \( P = P_{B,B'} \) and \( x = \sum_{i=1}^p \langle x, e_i \rangle e_i = \sum_{i=1}^{p-1} \langle x, e'_i \rangle e'_i \). The hypervolume is invariant by translation so we make the projection of \( S\ell_2^p \) onto \( \mathbb{R}^{p-1} \) to see that

\[ \text{Vol}(S\ell_2^p) = \text{Vol}(B_{2,p-1}(0, r_p)), \]

with \( r_p = 1/\sqrt{p(p-1)} \) the radius of the \( \ell_2 \) inscribed ball of \( \Delta_p \). This gives the value of \( \text{Vol}(S\ell_2^p) \).

**Mexican set.** Finally for the Mexican set, we cut off with a threshold \( \tau \) the length \( L = \sqrt{(p-1)/p} \) between the barycenter \( \bar{x} \) and a vertex \( e_i \). We get \( p \) smaller simplices and the volume we want is nothing but the difference between the volume of the simplex \( \Delta_p \) and \( p \) times the volume of a small simplex. To compute the hypervolume of one small simplex, we need to find its side length, knowing that its height is \( (1 - \tau)L \). We find a side length equal to

\[ \sqrt{2(1 - \tau)(p-1)/p} \]

and can conclude for the value \( \text{Vol}(S\ell_1^p) \).

We present in Figure 3 the evolution of the ratio \( \rho(S^\tau) \) of the Mexican set for different values of threshold and dimension.

![Figure 3: Evolution of \( \rho(S^\tau) \) with varying values of \((\tau, p)\).](image-url)
B Numerical experiments details

B.1 Additional results Feature Clustering

We present the full results of the performance of MEXICO regarding the feature clustering task. The projection step is either performed using alternating projections based on the method POCS (Projection Onto Convex Sets) or with the more elaborate technique Dykstra.

| \(p\) | Spherical-Kmeans \([33]\) | MEXICO (POCS) | MEXICO (Dykstra) |
|------|-----------------|---------------|-----------------|
|      | \(H\)           | \(C\)         | \(v-M\)         | \(H\)           | \(C\)         | \(v-M\)         | \(H\)           | \(C\)         | \(v-M\)         |
| 75   | 0.950±0.034     | 0.972±0.024   | 0.961±0.027     | 0.978±0.025     | 0.976±0.024   | 0.977±0.024     |
| 100  | 0.943±0.031     | 0.967±0.024   | 0.955±0.026     | 0.976±0.020     | 0.979±0.021   | 0.976±0.020     |
| 150  | 0.940±0.026     | 0.962±0.020   | 0.951±0.022     | 0.973±0.015     | 0.977±0.013   | 0.975±0.014     |
| 200  | 0.940±0.018     | 0.962±0.014   | 0.951±0.015     | 0.970±0.015     | 0.975±0.012   | 0.972±0.013     |

Table 4: Comparison of Homogeneity (\(H\)), Completeness (\(C\)) and v-Measure (\(v-M\)) from prediction scores for SphericalKmeans and Mexico with alternating projections on simulated data with different dimension \(p\).

| \(p\) | Spherical-Kmeans \([33]\) | MEXICO (Dykstra) |
|------|-----------------|-----------------|
|      | \(H\)           | \(C\)         | \(v-M\)         | \(H\)           | \(C\)         | \(v-M\)         |
| 75   | 0.950±0.034     | 0.972±0.024   | 0.961±0.027     | 0.977±0.025     | 0.976±0.024   | 0.977±0.024     |
| 100  | 0.943±0.031     | 0.967±0.024   | 0.955±0.026     | 0.978±0.020     | 0.979±0.021   | 0.978±0.020     |
| 150  | 0.940±0.026     | 0.962±0.020   | 0.951±0.022     | 0.976±0.015     | 0.980±0.013   | 0.978±0.014     |
| 200  | 0.940±0.018     | 0.962±0.014   | 0.951±0.015     | 0.967±0.015     | 0.972±0.012   | 0.970±0.013     |

Table 5: Comparison of Homogeneity (\(H\)), Completeness (\(C\)) and v-Measure (\(v-M\)) from prediction scores for SphericalKmeans and Mexico with Dykstra projection on simulated data with different dimension \(p\).

B.2 Anomaly detection, real world data preprocessing

We present the details about the preprocessing of the real world datasets.

The shuttle dataset is the fusion of the training and testing datasets available in the UCI repository \([37]\). The data have 9 numerical attributes, the first one being time. Labels from 7 different classes are also available. Class 1 instances are considered as normal, the others as anomalies. We use instances from all different classes but class 4, which yields an anomaly ratio (class 1) of 7.2%.

In the forestcover data, also available at UCI repository \([37]\), the normal data are the instances from class 2 while instances from class 4 are anomalies, other classes are omitted, so that the anomaly ratio for this dataset is 0.9%.

The last three datasets belong to the KDD Cup 99 dataset \([35, 50]\), produced by processing the tcpdump portions of the 1998 DARPA Intrusion Detection System (IDS) Evaluation dataset, created by MIT Lincoln Lab \([38]\). The artificial data was generated using a closed network and a wide variety of hand-injected attacks (anomalies) to produce a large number of different types of attack with normal activity in the background. Since the original demonstrative purpose of the dataset concerns supervised AD, the anomaly rate is very high (80%), which is unrealistic in practice, and inappropriate for evaluating the performance on realistic data. We thus take standard preprocessing steps in order to work with smaller anomaly rates.

For datasets SF and http we proceed as described in \([56]\): SF is obtained by picking up the data with positive logged-in attribute, and focusing on the intrusion attack, which gives an anomaly proportion of 4.5%. The dataset http is a subset of SF corresponding to a third feature equal to 'http'. Finally, the SA dataset is obtained as in \([26]\) by selecting all the normal data, together with a small proportion (3.4%) of anomalies.

We present the full results of the performance of MEXICO regarding the anomaly detection task. The projection step is either performed using alternating projections based on the method POCS (Projection Onto Convex Sets) or with the more elaborate technique Dykstra.
### Table 6: Comparison of Area Under Curve of Receiver Operating Characteristic (ROC-AUC) from prediction scores of each method on different anomaly detection datasets.

| Dataset     | iForest [39] | Damex [28] | Mexico (POCS) | Mexico (Dykstra) |
|-------------|--------------|------------|----------------|------------------|
| SF          | 0.381±0.086  | 0.710±0.031| **0.892±0.013**| 0.710±0.030      |
| SA          | 0.886±0.032  | 0.982±0.002| 0.981±0.006    | 0.983±0.031      |
| http        | 0.656±0.094  | 0.996±0.002| 0.995±0.005    | 0.997±0.002      |
| shuttle     | 0.970±0.020  | 0.990±0.003| 0.990±0.003    | 0.989±0.003      |
| forestcover | 0.654±0.096  | 0.762±0.008| **0.863±0.015**| 0.851±0.008      |

### Table 7: Comparison of Average Precision (AP) from prediction scores of each method on different anomaly detection datasets.

| Dataset     | iForest [39] | Damex [28] | Mexico (POCS) | Mexico (Dykstra) |
|-------------|--------------|------------|----------------|------------------|
| SF          | 0.393±0.081  | 0.650±0.034| **0.812±0.016**| 0.661±0.031      |
| SA          | 0.879±0.031  | 0.938±0.012| 0.940±0.031    | 0.950±0.011      |
| http        | 0.658±0.099  | 0.968±0.009| 0.972±0.012    | 0.971±0.008      |
| shuttle     | 0.825±0.055  | 0.864±0.026| 0.864±0.037    | 0.818±0.024      |
| forestcover | 0.894±0.037  | 0.893±0.010| 0.958±0.006    | 0.954±0.004      |

### C Further Numerical Experiments

The authors of [17] provide an archetypal analysis of the Swiss Army dataset. This dataset consists of 6 head dimensions from 200 Swiss soldiers. The data was gathered to construct face masks for the Swiss army. Few samples of the dataset are presented in Table 8.

The first measurement (MFB) corresponds to the width of the face just above the eyes. The second feature (BAM) corresponds to the width of the face just below the mouth. The third measurement (TFH) is the distance from the top of the nose to the chin. The fourth feature (LGAN) is the length of the nose. The fifth measurement (LTN) is the distance from the ear to the top of the head while the sixth (LTG) is the distance from the ear to the bottom of the face. For a better visualization of the dataset, we made simple drawings of the different samples. Figure 4a illustrates the 6 measurements.

| id | MFB | BAM  | TFH  | LGAN | LTN  | LTG  |
|----|-----|------|------|------|------|------|
| 0  | 113.2 | 111.7 | 119.6 | 53.9  | 127.4 | 143.6 |
| 1  | 117.6 | 117.3 | 121.2 | 47.7  | 124.7 | 143.9 |
| 2  | 112.3 | 124.7 | 131.6 | 56.7  | 123.4 | 149.3 |
| 3  | 116.2 | 110.5 | 114.2 | 57.9  | 121.6 | 140.9 |
| 4  | 112.9 | 111.3 | 114.3 | 51.5  | 119.9 | 133.5 |

Table 8: Extract of the Swiss Army dataset.

A question that naturally rises is to figure out subgroups of face features that get large simultaneously. Mexico algorithm performed on the standardized dataset provides the following groups of features: {5, 6} (green), {1, 3} (blue) and {2, 4} (red), as illustrated in Figure 4b.

Figure 4: Illustration of the 6 measurements (a) and subgroups that tend to be large simultaneously (b).