Feature clustering for extreme events analysis, with application to extreme stream-flow data

Maël Chiapino and Anne Sabourin

LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, Paris, France

Abstract. The dependence structure of extreme events of multivariate nature plays a special role for risk management applications, in particular in hydrology (flood risk). In a high dimensional context ($d > 50$), a natural first step is dimension reduction. Analyzing the tails of a dataset requires specific approaches: earlier works have proposed a definition of sparsity adapted for extremes, together with an algorithm detecting such a pattern under strong sparsity assumptions. Given a dataset that exhibits no clear sparsity pattern we propose a clustering algorithm allowing to group together the features that are 'dependent at extreme level', i. e., that are likely to take extreme values simultaneously. To bypass the computational issues that arise when it comes to dealing with possibly $O(2^d)$ subsets of features, our algorithm exploits the graphical structure stemming from the definition of the clusters, similarly to the Apriori algorithm, which reduces drastically the number of subsets to be screened. Results on simulated and real data show that our method allows a fast recovery of a meaningful summary of the dependence structure of extremes.

Keywords: extreme values; dimension reduction; pattern mining; subspace clustering; subgroup discovery

1 Introduction

Extreme value analysis is of primarily interest in many contexts. One example is the machine learning problem of anomaly detection, where one needs to control the false positive rate in the most remote regions of the sample space ([7,21,16,17]). Another example is the field of environmental sciences, where extreme events (floods, droughts, heavy rainfall, ...) are of particular concern to risk management, considering the disastrous impact these events may have. Using Extreme Value Theory (EVT) as a general setting to understand or predict extreme events has a long history ([20]). In spatial problems, exhibiting areas (groups of weather stations) which may be concomitantly impacted by severe events is of direct interest for risk management policies. Identifying these groups may also serve as a preliminary dimensionality reduction step before more precise modeling. Before proceeding further, we emphasize that standard dimension reduction techniques such as PCA do not apply to extremes as these methods essentially focus on the data around the mean by analyzing their covariance structure, which does

not characterize the behavior of extremes (i. e., data far away in the tails of the distribution). In the present paper, the quantity of interest is river water-flow recorded at several locations of the French river system. The features of the experiment are thus the stream-flow records at different gauging stations, and the goal is to recover maximal groups of stations where extreme discharge may occur simultaneously. Our dataset consists of daily stream-flow recorded at 92 gauging stations scattered over the French river system, from 1969, January 1^{st} to 2008, December 31^{st}. It is the same dataset as in [14], up to 220 gauging stations presenting missing or censored records, which have been removed from our analysis, which results in $n = 14610$ vectors $X_1, .., X_n$ in $\mathbb{R}^d$, with $d = 92$ the number of stations. The reader is referred to [14] for more details.

Related work. For the purpose of anomaly detection, [16,17] proposed a method to learn the sparsity pattern of the dependence structure of extremes: the aim is to recover the groups of components (features) which may take large values simultaneously, *while the other features stay small.* To the best of our knowledge, these works are the only ones that tackle this specific problematic. The output of [16,17]'s DAMEX algorithm is a (hopefully sparse) vector $\hat{\mathcal{M}} = (\hat{\mu}_\alpha, \alpha \subset \{1, \dots, d\})$ of size $2^d - 1$, where $\hat{\mu}_\alpha$ is a summary of the dependence strength at extreme levels between features $j \in \alpha$. The fact that $\hat{\mu}_\alpha$ is positive means that the probability that all features in α be large while all others stay small, is not negligible. Various datasets have been analyzed in [16,17] for which the DAMEX algorithm does exhibit a sparsity pattern, thus pointing to a relatively small number of groups of features α (each being of relatively small size $|\alpha|$ compared to the original dimension of the problem) which could be jointly extreme. However, DAMEX becomes unusable in situations where the subsets of features impacted by extreme events vary from one event to another: DAMEX then finds a very large number of subsets to be dependent, but not significantly so, (i. e., $0 < \hat{\mu}_\alpha \ll 1$), so that no sparsity pattern emerges. This is precisely the case with the river flow dataset analyzed in the present paper (see Section 5).

Contributions. One remarkable aspect of preliminary analysis of the river flow dataset using DAMEX is the tendency of those many subsets α's such that $\hat{\mu}_\alpha > 0$, to form *clusters*, whose members differ from each other by a single or two features only. In practice, this means that several distinct events have impacted 'almost' the same group (cluster) of stations. The aim of this paper is to propose a methodology enabling to gather together such 'close-by' feature subsets into feature clusters. This is done by relaxing the constraint that 'features not in α take small value' when constructing the representation of the dependence structure. The output of the CLEF algorithm (CLustering Extreme Features) proposed in the present work (Section 4) is an alternative representation which remains usable in this 'weakly sparse' context. This representation can still be explained and understood in the multivariate EVT framework (Section 3), as in [16,17].

Relationships with Apriori This dimension reduction framework (determining which subgroups of features are dependent at extreme level) is closely related to the problem of frequent itemsets mining, specifically to the well known Apriori

algorithm introduced by [2], see also [19]. The combinatorial issue that arises with possibly $2^d - 1$ subsets is circumvented in Apriori by considering subsets of increasing sizes, letting a subset 'grow' until its frequency in the database is not significant anymore. This incremental principle is also related to a subset clustering method proposed in [1]. CLEF proceeds in a similar way to Apriori, the main difference being the stopping criterion used to decide whether incrementing a feature subset, and the fact that the output has a natural interpretation in the framework of multivariate EVT.

The paper is organized as follows. Section 2 sets up the extremal feature clustering problem and establishes connections with multivariate EVT. The dimension reduction device is explained in Section 3. Section 3.1 recalls existing work and points out some limitations, Section 3.2 makes explicit the links between the considered problem and the Apriori algorithm. The CLEF algorithm is described in Section 4. Section 5 gathers results: the output of CLEF is compared with that of DAMEX and Apriori. Section 6 concludes. The `Python` code for CLEF, the scripts and the dataset used for our hydrological case study are available at https://bitbucket.org/mchiapino/clef_algo.

2 Problem statement and multivariate EVT viewpoint

2.1 Formal statement of the problem

Consider a multivariate random quantity of interest $X = (X^1, \dots, X^d)$ in $\mathbb{R}^d$ (here, X^j is the water discharge recorded at location j). The first step when it comes to learning dependence properties of X is to standardize the features, in the same spirit as in the copula framework; and one possible choice for that is the probability integral transform: Denote by F the joint cumulative distribution function (*c.d.f.*) of X and by F^j the marginal *c.d.f.* of X^j. Define $V^j = (1 - F^j(X^j))^{-1}$, and $V = (V^1, \dots, V^d)$, which allows us to focus only on the *dependence* structure of X. Our goal here is to recover all the maximal subsets of features (stations) $\alpha \subset \{1, \dots, d\}$ which 'may be large together' with non negligible probability. In more formal terms, define the *extremal joint excess coefficient*,

$$\rho_\alpha := \lim_{t\to\infty} t\mathbb{P}\left(\forall j \in \alpha, V^j > t\right) = \lim_{t\to\infty} \mathbb{P}\left(\forall j \in \alpha, V^j > t \mid V^{\alpha_1} > t\right) \in [0,1] \quad (1)$$

Such a limit exists under the regularity property (3) in the next paragraph. Notice already that the second equality comes from our standardization choice: if F^j is continuous, then for any $j \leq d$, $t^{-1} = \mathbb{P}(V^j > t) = \mathbb{P}(V^{\alpha_1} > t)$, which justifies the scaling factor t in the definition. The coefficient $\rho_\alpha \in [0,1]$ may be seen as a 'correlation' coefficient for the features $X^j, j \in \alpha$ at extreme levels. We say that the features $\{V^j, j \in \alpha\}$ 'may be large together' if $\rho_\alpha > 0$. One relevant summary of the dependence structure of extremes is thus the set of subgroups

$$\mathbb{M} = \{\alpha \subset \{1, \dots, d\} \ : \quad \rho_\alpha > 0\}. \quad (2)$$

More precisely, we would like to recover those subgroups $\alpha \in \mathbb{M}$ which are maximal for inclusion in $\mathbb{M}$, i. e., $\forall \beta$ such that $\alpha \subsetneq \beta,\ \beta \notin \mathbb{M}$. A maximal set of features $\alpha \in \mathbb{M}$ may be viewed as a *cluster*, in the sense that every subset $\beta \subset \alpha$ is dependent at extreme level (i. e., $\rho_\beta > 0$), and that α 'gathers' all of them together. In this paper, a 'cluster' of features is understood as a maximal element $\alpha \in \mathbb{M}$.

2.2 Connections with multivariate EVT

The working hypothesis in EVT is that, up to marginal standardization, the distribution of X is 'approximately homogeneous' on extreme regions. As pointed out above, if the margins F^j are continuous, then the V^j's have the homogeneity property: $t\mathbb{P}\left(\frac{V^j}{t} \geq x\right) = 1/x$, for $1 \leq j \leq d$, $t > 1$, $x > 0$. The key assumption is that the latter property holds *jointly* at extreme levels, i. e., that V is jointly *regularly varying* (see *e.g.* [23]), which writes

$$t\mathbb{P}\left(\frac{V}{t} \in A\right) \xrightarrow[t\to\infty]{} \mu(A), \tag{3}$$

where μ is the so-called *exponent measure* and where A is any set in $\mathbb{R}^d$ which is bounded away from 0 and such that $\mu(\partial A) = 0$. The exponent measure is finite on any such set A and satisfies, for $t > 0, A \subset \mathbb{R}^d_+$, $t\mu(tA) = \mu(A)$, where $tA = \{tx : x \in A\}$. Notice that many commonly used textbook multivariate distributions (*e.g.* multivariate Gaussian or Student distributions) satisfy (3), after standardization to V variables. The measure μ characterizes the distribution of V at extreme levels, since for t large enough (so that the region tA is an 'extreme region' of interest), one may use the approximation $\mathbb{P}(V \in tA) \simeq t^{-1}\mu(A)$. The connection between μ and the ρ_α's is as follows: consider the 'rectangle'

$$\Gamma_\alpha := \{x \in \mathbb{R}^d_+ :\ \ \forall j \in \alpha,\ x^j > 1\} \tag{4}$$

From the definitions (1) and (3), it follows that $\rho_\alpha = \mu(\Gamma_\alpha)$. Thus the family of subset $\mathbb{M}$ in (2) writes $\mathbb{M} = \{\alpha : \mu(\Gamma_\alpha) > 0\}$.

Non parametric estimation In a word, non parametric estimation of extremal characteristics based on *i.i.d.* data $X_1, \ldots, X_n$ (distributed as X) is performed by replacing probability distributions with their empirical counterparts, and by proceeding as if the limit in (3) were reached above some large, fixed threshold t, which is chosen depending on the sample size n. Theoretical guarantees on the estimators are obtained for $t = n/k$ where $k = o(n)$ and $k \to \infty$ (typically, $k \approx \sqrt{n}$, see *e.g.* [3], Chapter 3 for more details). Since the F^j's are unknown, set $\hat{V}_i^j = \frac{1}{1-\hat{F}^j(X_i^j)},\ i = 1, \ldots, n,\ j = 1, \ldots, d$, where $\hat{F}^j(\,\cdot\,)$ is an empirical version of the cumulative distribution function. Then the exponent measure μ of any region $A \subset \mathbb{R}^d_+ \setminus \{0\}$ is approximated by

$$\mu_n(A) = t\hat{P}_n(tA), \qquad \text{where } \ \hat{P}_n(A) = n^{-1}\sum_{i=1}^{n} \delta_{\hat{V}_i}(A). \tag{5}$$

Statistical properties of μ_n (or of other functional summaries of it) have been investigated by many authors, see *e.g.* [22,11,12] for the asymptotic behavior, [15] for finite sample error bounds.

3 Dimension reduction for multivariate extremes

3.1 Existing work

Numerous modeling strategies for extremes of moderate dimension (say $d \leq 10$) have been proposed, see *e.g.* [9,10,25] for parametric modeling, [4,18,24,13] for semi- or non-parametric ones. In order to address higher dimensional problems, dimensionality reduction algorithms have recently been proposed ([6,16,17]). The latter references share the common idea of recovering the sub-cones of $\mathbb{R}_+^d$ on which the exponent measure μ concentrates. The present work is mainly related to [16,17] insofar as we restrict the search to a finite number of regions that are defined by constraints 'parallel to the axes', as it is the case in (4). [16,17] consider the truncated cones

$$\mathcal{C}_\alpha = \{x : \|x\|_\infty \geq 1, x_j > 0 \text{ for } j \in \alpha \ ; \ x_j = 0 \text{ for } j \notin \alpha\}. \tag{6}$$

The importance of such cones in the analysis comes from the homogeneity property of μ. More precisely, a subset of features α may take large values together while the other take small values, if and only if μ assigns a positive mass to $\mathcal{C}_\alpha$. The approach proposed in [17] consists in 'thickening' the cones $\mathcal{C}_\alpha$, i. e., defining for some small $\epsilon > 0$ (typically, $\epsilon = 0.1$),

$$\mathcal{C}_{\alpha,\epsilon} = \{x \in \mathbb{R}_+^d : \ \|x\|_\infty \geq 1 \ ; \ \|x\|_\infty^{-1} x_j > \epsilon \text{ for } j \in \alpha \ ; \ \|x\|_\infty^{-1} x_j \leq \epsilon \text{ for } j \notin \alpha\}. \tag{7}$$

The quantity $\mu_\alpha := \mu(\mathcal{C}_\alpha)$ is approximated by its empirical counterpart on $\mathcal{C}_{\alpha,\epsilon}$, $\hat{\mu}(\mathcal{C}_\alpha) = \mu_n(\mathcal{C}_{\alpha,\epsilon})$, where μ_n is the empirical estimator defined in (5). In practice a tolerance parameter $\mu_{\min}$ has to be chosen: for any α such that $\mu_n(\mathcal{C}_{\alpha,\epsilon}) < \mu_{\min}$, one sets $\hat{\mu}(\mathcal{C}_\alpha) = 0$. The final output of [17]'s DAMEX algorithm is the potentially sparse $2^d - 1$-vector $\hat{\mathcal{M}} = (\hat{\mu}_\alpha)_{\alpha \subset \{1,\ldots,d\}}$ mentioned in the introduction, with $\hat{\mu}_\alpha := \hat{\mu}(\mathcal{C}_\alpha)$.

One shortcoming of DAMEX is that no sparsity pattern is produced in case of 'noise', i. e., when the empirical extreme mass is spread over many sub-cones $\mathcal{C}_{\alpha,\epsilon}$'s. This suggests an alternative approach allowing to gather together those α's that are 'close', where being 'close' means belonging to a single relevant super-set.

3.2 Gathering together 'close-by' cones

One way to 'gather' different $\mathcal{C}_{\alpha,\epsilon}$'s is to relax the condition that 'all the features V^j for $j \notin \alpha$ take small values' in the definition of $\mathcal{C}_{\alpha,\epsilon}$. This yields the rectangular region Γ_α defined in (4). Unlike the regions $\mathcal{C}_{\alpha,\epsilon}$'s, the Γ_α's do not form a partition of the positive orthant of $\mathbb{R}^d$, and indeed the fact that a point V_i belongs to Γ_α

does not tell anything about its features V_i^j for $j \notin \alpha$. The problem addressed in [17] (recovering $\mathcal{M} := \{\alpha : \mu(\mathcal{C}_\alpha) > 0\}$) and the relaxed problem considered here (recovering $\mathbb{M} := \{\alpha : \rho_\alpha > 0\} = \{\alpha : \mu(\Gamma_\alpha) > 0\}$) are different but however related through the maximal elements of $\mathcal{M}$ and $\mathbb{M}$, as stated in the following lemma. Recall that α is said to be maximal in $\mathbb{M}$ (*resp.* $\mathcal{M}$) if there is no superset $\alpha' \supsetneq \alpha$ in $\mathbb{M}$ (*resp.* $\mathcal{M}$).

Lemma 1. *For $\alpha \subset \{1, \dots, d\}$,*

$$\alpha \text{ is maximal in } \mathbb{M} \Leftrightarrow \alpha \text{ is maximal in } \mathcal{M}. \tag{8}$$

The proof is deferred to the Appendix.
Another important property from an algorithmic perspective is the following:

Lemma 2. *For $\alpha \subset \{1, \dots, d\}$, if $\rho_\alpha = 0$ then also for all $\alpha' \supset \alpha$, $\rho_{\alpha'} = 0$.*

The proof is immediate: remind that $\rho_\alpha = \mu(\Gamma_\alpha)$ and notice that for $\alpha \subset \alpha'$, $\Gamma_{\alpha'} \subset \Gamma_\alpha$.

Lemma 2 suggests an incremental-type algorithm such as Apriori ([2]) meaning that one may search for α's such that $\rho_\alpha > 0$ among α's of increasing size following the Hasse diagram, and stopping the search along a given path of the latter diagram as soon as $\rho_\alpha = 0$ for some α. This incremental strategy is the main ingredient of the Apriori algorithm, which we recall for convenience: Let $I = \{item_1, \dots, item_d\}$ be a set of items and let $T = \{t_1, \dots, t_n\}$ be a set of transactions with $t_i \subset I, \forall i \in \{1, \dots, n\}$. The frequency of appearance of the list of items $\alpha \subset I$ is defined as $f_\alpha := \frac{1}{n} \sum_{1 \leq i \leq n} \mathbb{1}_{\alpha \subset t_i}$. Apriori returns the set $\{\alpha : f_\alpha > f_{min}\}$ with $f_{min} > 0$. It begins with pairs of items and then increases the size of the subsets at each step. Indeed if $f_\alpha \leq f_{min}$ then all supersets $\alpha' \supset \alpha$ verify $f_{\alpha'} \leq f_{min}$ as well, which reduces drastically the number of subsets to be tested.

4 Empirical criterion and implementation

4.1 Conditional criterion for extremal dependence

Considering the relaxed framework where the goal is to recover the set $\mathbb{M}$ defined in (2), one needs an empirical criterion allowing to test the condition $\rho_\alpha (= \mu(\Gamma_\alpha)) > 0$. One option would be to consider the empirical estimator $\hat{\rho}_\alpha = \mu_n(\Gamma_\alpha)$ where μ_n is defined in (5), then to set $\hat{\rho}_\alpha = 0$ whenever $\hat{\rho}_\alpha \leq \rho_{\min}$, with $\rho_{\min}$ a user-defined tolerance level. However, since the Γ_α's (for increasing α's) are nested, the quantities $\mu(\Gamma_\alpha)'s$ are decreasing with the size of α. The threshold $\rho_{\min}$ should thus depend on the size of α, which would introduce d tuning parameters instead of one. The alternative considered in the present paper is to compare $\mu_n(\Gamma_\alpha)$ with $\mu_n(\Gamma_\beta)$, with $\beta \subset \alpha$. More precisely, consider the probability that all the features in α be large given that all of them but at most one are large. This yields the conditional coefficient:

$$\kappa_\alpha := \frac{\mu(\Gamma_\alpha)}{\mu(\Delta_\alpha)} \tag{9}$$

with $\Delta_\alpha := \{x \in \mathbb{R}_+^d : \|x\|_\infty > 1, \sum_{j\in\alpha} \mathbb{1}_{x_j \geq 1} \geq |\alpha| - 1\}$. The idea is now to compare empirical counterparts of κ_α with a single fixed tolerance parameter $\kappa_{\min} > 0$. This amounts to decide that $\mu_n(\Gamma_\alpha)$ results from noise if $\mu_n(\Gamma_\alpha) \ll \mu_n(\Delta_\alpha)$. Notice that $\Gamma_\alpha \subset \Delta_\alpha$, so that $\kappa_\alpha \in [0, 1]$ whenever the denominator in (9) is well defined. This is the case if and only if $\mu_n(\Gamma_\beta) > 0$ for some $\beta \subset \alpha$ such that $|\alpha \setminus \beta| = 1$, which is another argument in favor of an incremental strategy.

4.2 Algorithm

CLEF (summarized in Algorithm 1) uses the empirical counterpart of the conditional criterion κ_α, which depends on a (high) threshold t as in (5):

$$\hat{\kappa}_{\alpha,t} := \frac{\mu_n(\Gamma_\alpha)}{\mu_n(\Delta_\alpha)} = \frac{\sum_{i=1}^n \mathbb{1}_{\left\{\#\{j\in\alpha:\ \hat{V}_i^j > t\} = |\alpha|\right\}}}{\sum_{i=1}^n \mathbb{1}_{\left\{\#\{j\in\alpha:\ \hat{V}_i^j > t\} \geq |\alpha|-1\right\}}}. \tag{10}$$

For $k \geq 2$, families $\hat{\mathcal{A}}_k$ of subsets α of size k are constructed in an incremental way, among a set of candidates $\mathcal{A}'_k$, as follows: Set $\hat{\mathcal{A}}_1 = \{\{1\}, \ldots, \{d\}\}$, then

$$\begin{aligned} \mathcal{A}'_k &= \Big\{\alpha \subset \{1, \ldots, d\} : |\alpha| = k, \forall \beta \subset \alpha \text{ s.t. } |\beta| = k-1 : \beta \in \hat{\mathcal{A}}_{k-1}\} \\ \hat{\mathcal{A}}_k &= \{\alpha \in \mathcal{A}'_k : \hat{\kappa}_{\alpha,t} > \kappa_{\min}\Big\}. \end{aligned} \tag{11}$$

Remark 1 (Choice of the parameters t and $\kappa_{\min}$). According to standard good practice in EVT (see *e.g. [8]*), t and $\kappa_{\min}$ are chosen in 'stability regions' of relevant summaries of the output. Here we consider the cardinal of $\hat{\mathbb{M}}$ and the mean cardinal of maximal subsets $\alpha \in \hat{\mathbb{M}}$, and t is chosen such that, when slightly increased, the output remains stable.

The procedure stops at step $K \leq d-1$ if $\hat{\mathcal{A}}_{K+1} = \emptyset$, at which point our estimator of the family $\mathbb{M}$ of dependent subsets is $\hat{\mathbb{M}} = \cup_{k=1}^K \hat{\mathcal{A}}_k$. Notice that restricting the search to the set of candidates $\mathcal{A}'_k$ ensures that the 'empirical counterpart' of Lemma 2 is satisfied, namely $\alpha \notin \hat{\mathbb{M}} \Rightarrow \forall \beta \supset \alpha, \beta \notin \hat{\mathbb{M}}$. It also avoids division by zero when computing (11). The final output of CLEF is the set $\hat{\mathbb{M}}_{\max}$ of maximal elements of $\hat{\mathbb{M}}$.

Remark 2. [Construction of the candidates $\mathcal{A}'_{k+1}$] The graphical structure of the patterns (subsets) is exploited as in the max-clique algorithm ([29]). Namely, members of $\mathcal{A}'_{k+1}$ are the maximal cliques of size k in the graph $(\mathcal{A}_k, \mathcal{E}_k)$, where $\mathcal{E}_k = \{(\alpha, \alpha') \in \mathcal{A}_k \times \mathcal{A}_k : |\alpha \cap \alpha'| = k-1\}$. Clique extraction is performed using Bron & Kerbosch ([5],[28]) algorithm, as implemented in the function `find_clique` of the `Python` package `NetworkX`.

Algorithm 1 CLEF (CLustering Extreme Features)

INPUT: High threshold t, tolerance parameter $\kappa_{\min} > 0$.
STAGE 1: constructing the $\hat{\mathcal{A}}_k$'s .
Initialization: set $K = d$.
Step 1: Construct the family of extremal-dependent pairs:
set $\hat{\mathcal{A}}_2 = \big\{\{i,j\} \subset \{1,\dots,d\}, \text{ such that } \hat{\kappa}_{\{i,j\}} > \kappa_{min}\big\}$.
Step 2: If $\hat{\mathcal{A}}_2 = \emptyset$, set $K = 2$; end **STAGE 1**. Otherwise

- generate candidate triplets $\mathcal{A}'_3 = \{i,j,k\} \subset \{1,\dots,d\}$ s.t $\{i,j\},\{i,k\},\{j,k\} \in \hat{\mathcal{A}}_2\}$,
- set $\hat{\mathcal{A}}_3 = \big\{\alpha \in \mathcal{A}'_3 \text{ s.t. } \hat{\kappa}_\alpha > \kappa_{min}\big\}$.

$\vdots$

Step k($k \leq d$): If $\hat{\mathcal{A}}_k = \emptyset$, set $K = k$; end **STAGE 1**. Otherwise

- generate candidates of size $k+1$, $\mathcal{A}'_{k+1} = \{\alpha \subset \{1,\dots,d\}, |\alpha| = k+1, \alpha \setminus \{j\} \in \hat{\mathcal{A}}_k \text{ for all } j \in \alpha\}$,
- set $\hat{\mathcal{A}}_{k+1} = \big\{\alpha \in \mathcal{A}'_{k+1} \text{ such that } \hat{\kappa}_\alpha > \kappa_{min}\big\}$.

Output: $\hat{\mathbb{M}} = \cup_{k=1}^{K} \hat{\mathcal{A}}_k$.
STAGE 2: pruning (keeping maximal α's only)
Initialization: $\hat{\mathbb{M}}_{\max} \leftarrow \hat{\mathcal{A}}_K$.
for $k = (K-1) : 2$, for $\alpha \in \hat{\mathcal{A}}_k$,
If there is no $\beta \in \hat{\mathbb{M}}_{\max}$ such that $\alpha \subset \beta$, $\hat{\mathbb{M}}_{\max} \leftarrow \hat{\mathbb{M}}_{\max} \cup \{\alpha\}$.
Output: $\hat{\mathbb{M}}_{\max}$

5 Results

5.1 Stream-flow data

The output of CLEF for the stream-flow data may be visualized in Figure 1 (Execution time: 0.09 s on recent 4 cores laptop computer). Following the heuristic mentioned in Remark 1, the extremal threshold t was fixed to 600, yielding $n = 202$ extreme events (time indexes i such that $\|\hat{V}_i\|_\infty \geq t$). The parameter $\kappa_{\min}$ was fixed to 0.3. A total number of 69 clusters (elements of $\hat{\mathbb{M}}_{\max}$) are returned by the CLEF algorithm, the size of which varies between 2 and 6. At first inspection, Figure 1 agrees with general climatologic facts: in the north-western part of France, the climate is driven by large scale oceanographic perturbations, so that extreme floods tend to impact a large number of gauging stations simultaneously. The south-eastern part of France is rather subject to localized events (*e.g.* the so-called 'orages Cévenols' in the vicinity of the Mediterranean coast). This yields smaller clusters, both in terms of number of stations and of spatial extent. As a comparison, Table 1 shows the outcome of [17]'s DAMEX algorithm with the stream-flow data. These results show that no matter the choice of the thickening parameter ϵ in (7), the data do not concentrate on 'a few' thickened cones $\mathcal{C}_{\alpha,\epsilon}$, instead most of the empirical mass is spread onto many of them. In other words, there are too many subcones with positive mass, but not in a significant way.

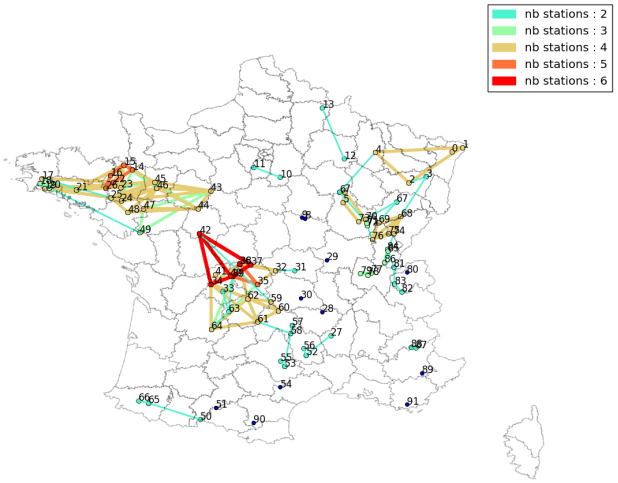

Fig. 1. Output of CLEF for the stream-flow dataset: Maximal groups of stations $\alpha \in \hat{\mathbb{M}}$ that are likely to be jointly impacted by an extreme event. Clusters of stations are marked by colored edges between their members, the color scale indicates the number of stations forming the cluster.

5.2 Simulation experiments

In order to assess the performance of CLEF in a supervised setting, we generate d-dimensional datasets under a model such that the exponent measure μ concentrates on K specified cones $(\mathcal{C}_{\alpha_1}, \ldots, \mathcal{C}_{\alpha_K})$. The generated data are 'realistic' in the sense that all the features are positive (the points lie in the interior cone $\mathcal{C}_{\{1,\ldots,d\}}$), even though the furthest points in the tails concentrate near the subcones $\mathcal{C}_{\alpha_k}$'s. Namely, we use the *asymmetric logistic* extreme value model ([27]), from which data is simulated using Algorithm 2.2 in [26]. 20 datasets of size $n = 100.10^3, d = 100$, are generated. For each dataset, K subsets $\alpha_1, \ldots, \alpha_K$ of $\{1, \ldots, d\}$ are randomly chosen, which sizes follow a truncated geometric distribution (the maximum cluster size is 8). We aim at reproducing the fact that different events associated with a single α usually impact a group of stations which differs from α by a few stations (the impacted area is not deterministic). To this end, for each step $i = 1, \ldots, n$, and each subset α_j, $j = 1, \ldots, K$, one randomly chosen 'noisy' feature $l_{i,j} \in \{1, \ldots, d\} \setminus \alpha_j$ is added to α_j. For CLEF, DAMEX and Apriori algorithms, the extreme threshold parameter t is chosen so that $\frac{\#\{i \leq n : \|\hat{V}_i\|_\infty \geq t\}}{n} \approx 5\%$. Table 2 summarizes the average performance of the two algorithms, for $K = 40, 50, 60, 70$. In these experiments, the CLEF algorithm recovers most of the charged K subsets $\alpha_1, \ldots, \alpha_K$ in average, and significantly more than Apriori. It should be noted though, that in situations (not reported here) where no noise is added, Apriori performs better than CLEF. As expected, in our 'noisy' simulations, DAMEX does not recover the sparse structure of the data.

Table 1. Output of [17]'s DAMEX algorithm with the hydrological dataset. Columns 1 and 2 indicate respectively the number of thickened cones $\mathcal{C}_{\alpha,\epsilon}$ with non zero empirical mass, and the percentage of cones (among those such that $\mu_n(\mathcal{C}_{\alpha,\epsilon}) > 0$) containing less than 1% of the 'extreme data', that is of $\#\{i : \|\hat{V}_i\|_\infty > t\}$.

ϵ	$\#\ \{\alpha : \mu_n(\mathcal{C}_{\alpha,\epsilon}) > 0\}$	$\%\left\{\alpha : \frac{\#\{i : t^{-1}V_i \in \mathcal{C}_{\alpha,\epsilon}\}}{\#\{i : \|V_i\| \geq t\}} < 1\%\right\}$
0.01	740	100%
0.05	688	98%
0.1	639	94%
0.2	559	88%

Table 2. Average number of errors (non recovered and falsely discovered clusters) of CLEF, Apriori and DAMEX with simulated, noisy data.

K	# errors CLEF	# errors Apriori	# errors DAMEX
40	1.2	6.4	72.2
50	3.5	10.9	91.0
60	6.3	14.6	112.4
70	10.1	25.8	134.0

6 Conclusion

We propose a novel dimension reduction method for the analysis of extremes of multivariate datasets *via* feature clustering. This is done in adequacy with the framework of multivariate extreme value theory. The proposed algorithm makes use of the graphical structure of the problem, scanning the multiple possible subsets of features in a time efficient way. Results on a hydrological streamflow data and on simulated data demonstrate the relevance of this approach on datasets which would not exhibit any sufficiently sparse structure when analyzed with existing algorithms. Future work will focus on the statistical properties of the empirical criteria $\hat{\kappa}_{\alpha,t}$ involved in the algorithm, which would allow to analyze the output as a statistical test for independence at extreme levels.

References

1. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data. DMKD 11(1), 5–33 (2005)
2. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB. vol. 1215, pp. 487–499 (1994)
3. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.: Statistics of extremes: theory and applications. John Wiley & Sons (2006)
4. Boldi, M.O., Davison, A.: A mixture model for multivariate extremes. JRSS-B 69(2), 217–229 (2007)
5. Bron, C., Kerbosch, J.: Algorithm 457: Finding all cliques of an undirected graph. Commun. ACM 16(9), 575–577 (Sep 1973)

6. Chautru, E., et al.: Dimension reduction in multivariate extreme value analysis. Electronic Journal of Statistics 9(1), 383–418 (2015)
7. Clifton, D.A., Hugueny, S., Tarassenko, L.: Novelty detection with multivariate extreme value statistics. Journal of signal processing systems 65(3), 371–389 (2011)
8. Coles, S.: An introduction to statistical modeling of extreme values. Springer Series in Statistics, Springer-Verlag, London (2001)
9. Coles, S., Tawn, J.: Modeling extreme multivariate events. JRSS-B 53, 377–392 (1991)
10. Cooley, D., Davis, R., Naveau, P.: The pairwise beta distribution: A flexible parametric multivariate model for extremes. JMVA 101(9), 2103–2117 (2010)
11. Einmahl, J.H., Segers, J.: Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. The Annals of Stat. pp. 2953–2989 (2009)
12. Fougeres, A.L., De Haan, L., Mercadier, C., et al.: Bias correction in multivariate extremes. The Annals of Stat. 43(2), 903–934 (2015)
13. Fougeres, A.L., Mercadier, C., Nolan, J.P.: Dense classes of multivariate extreme value distributions. Journal of Multivariate Analysis 116, 109–129 (2013)
14. Giuntoli, I., Renard, B., Vidal, J.P., Bard, A.: Low flows in france and their relationship to large-scale climate indices. J. of Hydro. 482, 105–118 (2013)
15. Goix, N., Sabourin, A., Clémençon, S.: Learning the dependence structure of rare events: a non-asymptotic study. In: Proc. of the 28th COLT (2015)
16. Goix, N., Sabourin, A., Clémençon, S.: Sparsity in multivariate extremes with applications to anomaly detection. arXiv preprint arXiv:1507.05899 (2015)
17. Goix, N., Sabourin, A., Clémençon, S.: Sparse representation of multivariate extremes with applications to anomaly ranking. In: Proceedings of the 19th AISTAT conference. pp. 287–295 (2016)
18. Guillotte, S., Perron, F., Segers, J.: Non-parametric bayesian inference on bivariate extremes. JRSS-B 73(3), 377–406 (2011)
19. Gunopulos, D., Khardon, R., Mannila, H., Saluja, S., Toivonen, H., Sharma, R.S.: Discovering all most specific sentences. ACM Trans. Database Syst. 28(2), 140–174 (Jun 2003)
20. Katz, R.W., Parlange, M.B., Naveau, P.: Statistics of extremes in hydrology. Advances in water resources 25(8), 1287–1304 (2002)
21. Lee, H.j., Roberts, S.J.: On-line novelty detection using the kalman filter and extreme value theory. In: Pattern Recognition. ICPR 2008. 19th International Conference on. pp. 1–4. IEEE (2008)
22. Qi, Y.: Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics. Acta Mathematicae Applicatae Sinica (English series) 13(2), 167–175 (1997)
23. Resnick, S.I.: Extreme values, regular variation and point processes. Springer (2013)
24. Sabourin, A., Naveau, P.: Bayesian dirichlet mixture model for multivariate extremes: A re-parametrization. CSDA 71, 542–567 (2014)
25. Sabourin, A., Naveau, P., Fougeres, A.L.: Bayesian model averaging for multivariate extremes. Extremes 16(3), 325 (2013)
26. Stephenson, A.: Simulating multivariate extreme value distributions of logistic type. Extremes 6(1), 49–59 (2003)
27. Tawn, J.A.: Modelling multivariate extreme value distributions. Biometrika 77(2), 245–253 (1990)
28. Tomita, E., Tanaka, A., Takahashi, H.: The worst-case time complexity for generating all maximal cliques and computational experiments. Theoretical Computer Science 363(1), 28–42 (2006)

29. Xie, Y., Philip, S.Y.: Max-clique: a top-down graph-based approach to frequent pattern mining. In: 2010 IEEE Int. Conf. Data Mining. pp. 1139–1144. IEEE (2010)

A Appendix: Proof of Lemma 1

Step 1. As a first step we show that $\mathcal{M} \subset \mathbb{M}$, i. e., $\mu(\mathcal{C}_\alpha) > 0 \Rightarrow \mu(\Gamma_\alpha) > 0$.

Proof. Write $\mathcal{C}_\alpha = \bigcup_{\epsilon>0, \epsilon\in\mathbb{Q}} R_{\alpha,\epsilon}$, where $R_{\alpha,\epsilon} = \{x \in \mathbb{R}_+^d : \|x\|_\infty \geq 1; \quad x_j > \epsilon \ (j \in \alpha); \quad x_i = 0 \ (i \notin \alpha)\}$. Assume $\mu(\mathcal{C}_\alpha) > 0$. Since $\mu(\mathcal{C}_\alpha) < \infty$, by the monotonous limit property of the measure μ, we have $\mu(\mathcal{C}_\alpha) = \lim_{\epsilon\to 0} \mu(R_{\alpha,\epsilon})$. Also, from the definitions, $R_{\alpha,\epsilon} \subset \epsilon\Gamma_\alpha$. Thus,

$$\begin{aligned}\mu(\mathcal{C}_\alpha) > 0 \Rightarrow \exists \epsilon \in (0,1) : \mu(R_{\alpha,\epsilon}) > 0 \quad &\Rightarrow \mu(\epsilon\Gamma_\alpha) > 0 \\ &\Rightarrow \rho_\alpha = \mu(\Gamma_\alpha) = \epsilon\mu(\epsilon\Gamma_\alpha) > 0.\end{aligned}$$

Step 2. We now prove the reverse inclusion for maximal elements of $\mathbb{M}$, i. e.,

$$\alpha \text{ is maximal in } \mathbb{M} \quad \Rightarrow \alpha \in \mathcal{M}. \tag{12}$$

Proof. Consider, for $i \notin \alpha$, the set $\Delta_{i,\epsilon} = \Gamma_\alpha \cap \{x \in \mathbb{R}_+^d : \quad x_i > \epsilon\}$, so that $\Gamma_\alpha = \Big\{ \bigcup_{\substack{i\in\{1,\dots,d\}\setminus\alpha \\ \epsilon\in\mathbb{Q}\cap(0,1)}} \Delta_{i,\epsilon} \Big\} \cup R_{\alpha,1}$. Thus,

$$\alpha \in \mathbb{M} \quad \Rightarrow \quad \mu(\Gamma_\alpha) > 0 \quad \Rightarrow \quad \Big(\exists i, \mu(\Delta_{i,\epsilon}) > 0 \quad \text{or} \quad \mu(R_{\alpha,1}) > 0\Big) \tag{13}$$

To prove (12), it is enough to show that

$$\alpha \in \mathbb{M} \quad \Rightarrow \quad \text{for } i \notin \alpha, \ \mu(\Delta_{i,\epsilon}) = 0. \tag{14}$$

Indeed if (14) is true, and if $\alpha \in \mathbb{M}$, then (13) implies that $\mu(R_{\alpha,1}) > 0$, and the result follows from the inclusion $R_{\alpha,1} \subset \mathcal{C}_\alpha$. We show (14) by contraposition. If $\mu(\Delta_{i,\epsilon}) > 0$ for some $i \notin \alpha$, then

$$\frac{1}{\epsilon}\Delta_{i,\epsilon} = \left(\frac{1}{\epsilon}\,\Gamma_\alpha\right) \cap \{x \in \mathbb{R}_+^d : x_i > 1\} \subset \Gamma_{\alpha\cup\{i\}},$$

thus $\mu(\Gamma_{\alpha\cup\{i\}}) > 0$, which contradicts the maximality of α in $\mathbb{M}$.

Step 3. From (12), if α is maximal in $\mathbb{M}$ then $\alpha \in \mathcal{M}$. Now if α is maximal in $\mathbb{M}$ but not in $\mathcal{M}$, there exists $\beta \supsetneq \alpha$ in $\mathcal{M}$. Thus from Step 1, $\beta \in \mathbb{M}$, a contradiction. Hence α is also maximal in $\mathcal{M}$. Conversely, if α is maximal in $\mathcal{M}$ then (Step 1) $\alpha \in \mathbb{M}$. If α was not maximal in $\mathbb{M}$, there would exists $\beta \supsetneq \alpha$ maximal in $\mathbb{M}$, and from (12), $\beta \in \mathcal{M}$, contradicting the maximality of α in $\mathcal{M}$.

Acknowledgments

Part of this work has been funded by the the 'LabEx Mathématiques Hadamard' (LMH) project, by the 'AGREED' project from the PEPS JCJC program (INS2I, CNRS) and by the chair 'Machine Learning for Big Data' from Télécom ParisTech. The authors would like to thank Benjamin Renard for interesting discussions about the hydrological use case and for sharing the data.