Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking

Nicolas Goix [*]
Institut Mines-Télécom
Télécom ParisTech
CNRS LTCI

Anne Sabourin
Institut Mines-Télécom
Télécom ParisTech
CNRS LTCI

Stéphan Clémençon
Institut Mines-Télécom
Télécom ParisTech
CNRS LTCI

Abstract

Capturing the dependence structure of multivariate extreme events is a major concern in many fields involving the management of risks stemming from multiple sources, *e.g.* portfolio monitoring, insurance, environmental risk management and anomaly detection. One convenient (nonparametric) characterization of extremal dependence in the framework of multivariate Extreme Value Theory (EVT) is the *angular measure*, which provides direct information about the probable 'directions' of extremes, that is, the relative contribution of each feature/coordinate of the 'largest' observations. Modeling the angular measure in high dimensional problems is a major challenge for the multivariate analysis of rare events. We propose here a novel methodology aiming at exhibiting a sparsity pattern within the dependence structure of extremes. This is done in a non-parametric way by estimating the amount of mass spread by the angular measure on representative sets of directions, corresponding to specific sub-cones of $\mathbb{R}_+^d$ (after non-linear transform of the data). This method performs linearly with the dimension and almost linearly in the data (in $O(dn \log n)$), and can thus be used for large scale problems. This dimension reduction technique paves the way towards scaling up existing multivariate EVT methods. Beyond a non-asymptotic study providing a theorical validity framework for our method, we propose as a direct application a –first– anomaly detection algorithm based on *multivariate* EVT. This algorithm builds a sparse 'normal profile' of extreme behaviours, to be confronted with new (possibly abnormal) extreme observations.

1 Context: multivariate extreme values in large dimension

Extreme value theory (EVT) provides a theorical basis for modeling the tails of probability distributions. Extremes are a central issue in many applied fields where rare events may have a disastrous impact, such as finance, insurance, climate, environmental risk management, network monitoring ([7, 14]), anomaly detection ([2, 10]). In a multivariate context, the dependence structure of the joint tail is of particular interest, as it gives access *e.g.* to probabilities of a joint excess above high thresholds or to multivariate quantile regions. Also, the distributional structure of extremes indicates which components of a multivariate quantity may be concomitantly large while the others are small, which is a valuable piece of information for multi-factor risk assessment or detection of anomalies among

[*] nicolas.goix@telecom-paristech.fr

other –not abnormal– extreme data. Indeed, extreme regions usually contain higher proportion of anomalies, and therefore play a very special role in anomaly detection.

Parametric or semi-parametric estimation of the structure of multivariate extremes is relatively well documented in the statistical literature, see *e.g.* [3, 8, 4, 13] and the references therein. In a multivariate 'Peak-Over-Threshold' setting, realizations of a d-dimensional random vector $\mathbf{V} = (V_1, ..., V_d)$ are observed and the goal pursued is to learn the conditional distribution of excesses, $[\, \mathbf{V} \mid \|\mathbf{V}\| \geq r \,]$, above some large threshold $r > 0$. The dependence structure of such excesses is described via the distribution of the 'directions' formed by the most extreme observations, the so-called *angular measure*, hereafter denoted by Φ. The latter is defined on the positive orthant of the $d-1$ dimensional hyper-sphere. To wit, for any region A on the unit sphere (a set of 'directions'), after suitable standardization of the data, $C\Phi(A) \simeq \mathbb{P}(\|\mathbf{V}\|^{-1}\mathbf{V} \in A \mid \|\mathbf{V}\| > r)$, where C is a normalizing constant. Some probability mass may be spread on any sub-sphere of dimension $k < d$, the k-faces of an hyper-cube if we use the infinity norm, which complexifies inference when d is large. To fix ideas, the presence of Φ-mass on a sub-sphere of the type $\{\max_{1\leq i\leq k} v_i = 1\,;\, v_i > 0\ (i \leq k)\,;\, v_{k+1} = \ldots v_d = 0\}$ indicates that the components $V_1, \ldots, V_k$ may simultaneously be large, while the others are small.

Scaling up multivariate EVT is a major challenge that one faces when confronted to high-dimensional learning tasks, since most multivariate extreme value models have been designed to handle moderate dimensional problems (say, of dimensionality $d \leq 10$). For larger dimensions, simplifying modeling choices are needed, stipulating *e.g* that only some pre-definite subgroups of components may be concomitantly extremes, or, on the contrary, that all of them must be (see *e.g.* [15] or [13]). 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. This calls for dimensionality reduction devices adapted to multivariate extreme values.

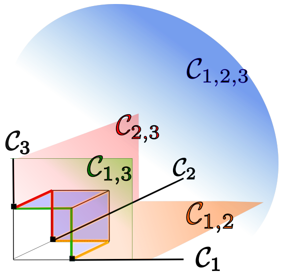

Figure 1: Truncated cones in 3D

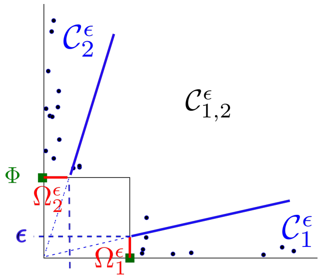

Figure 2: Truncated ϵ-cones in 2D

In a wide range of situations, one may expect the occurrence of two phenomena:

1- Only a 'small' number of groups of components may be concomitantly extreme, so that only a 'small' number of hyper-cubes (those corresponding to these subsets of indexes precisely) have non zero mass ('small' is relative to the total number of groups 2^d).

2- Each of these groups contains a limited number of coordinates (compared to the original dimensionality), so that the corresponding hyper-cubes with non zero mass have small dimension compared to d.

The main purpose of this work is to introduce a data-driven methodology for identifying such faces, so as to reduce the dimensionality of the problem and thus to learn a sparse representation of extreme behaviors. In case hypothesis **2-** is not fulfilled, such a sparse 'profile' can still be learned, but looses the low dimensional property of its supporting hyper-cubes. One major issue is that real data generally do not concentrate on sub-spaces of zero Lebesgue measure. This is circumvented by setting to zero any coordinate less than a threshold $\epsilon > 0$, so that the corresponding 'angle' is assigned to a lower-dimensional face.

More formally, Figures 1 and 2 represent the transformed input space, resulting from classical standardization of the margins. After this non-linear transform, the representation of extreme data is linear and learned by estimating the mass on the sub-cones

$$\mathcal{C}_\alpha = \{\mathbf{v} \geq 0,\ \|\mathbf{v}\|_\infty \geq 1,\ v_j > 0 \ \text{ for } j \in \alpha,\ v_j = 0 \ \text{ for } j \notin \alpha\},$$

or more precisely, the mass of the angular measure Φ on the corresponding sub-spheres

$$\Omega_\alpha = \{\mathbf{x} \in S_\infty^{d-1} : x_i > 0 \text{ for } i \in \alpha\ ,\ x_i = 0 \text{ for } i \notin \alpha\} = S_\infty^{d-1} \cap \mathcal{C}_\alpha,$$

represented in Figure 1. This is done using ϵ-thickened sub-cones $\mathcal{C}_\alpha^\epsilon$, corresponding to ϵ-thickened sub-spheres Ω_α^ϵ, as shown in Figure 2 in the two-dimensional case. We thus obtain an estimate $\widehat{\mathcal{M}}$ of the representation $\mathcal{M} = \{\Phi(\Omega_\alpha) : \ \emptyset \neq \alpha \subset \{1,\ \ldots,\ d\}\}$. Theoretically, recovering the $(2^d - 1)$-dimensional unknown vector $\mathcal{M}$ amounts to roughly approximating the support of Φ using the partition $\{\Omega_\alpha, \alpha \subset \{1, \ldots, d\}, \alpha \neq \emptyset\}$, that is, determine which Ω_α's have nonzero mass (and evaluating the mass $\Phi(\Omega_\alpha)$), or equivalently, Φ_α's are nonzero. This support estimation is potentially sparse (if a small number of Ω_α have non-zero mass, *i.e.* Phenomenon **1-**) and potentially low-dimensional (if the dimensions of the sub-spheres Ω_α with non-zero mass are low, *i.e.* Phenomenon **2-**).

2 Algorithm and Application to Anomaly Detection

The proposed algorithm, DAMEX for Detecting Anomalies with Extremes, was initially conceived for anomaly detection. DAMEX learns, as explained in Section 1, a (possibly sparse and low-dimensional) representation of the angular measure, rough enough to control the variance in high dimension and accurate enough to gain information about the 'probable directions' of extremes. From a theoretical perspective, it yields a –first– non-parametric estimate of the angular measure in any dimension, restricted to a class of sub-spheres.We believe that our method can also be used as a preprocessing stage, for dimensionality reduction purpose, before proceeding with a parametric or semi-parametric estimation of the angular measure which could benefit from the structural information issued in the first step. Such applications are beyond the scope of this work and will be the subject of further research.

EVT has been intensively used in anomaly detection in the one-dimensional situation, see for instance [11], [12], [2], [1], [10]. In the multivariate setup, however, there is –to the best of our knowledge– no anomaly detection method relying on *multivariate* EVT. Until now, the multidimensional case has only been tackled by means of extreme value statistics based on univariate EVT. The major reason is the difficulty to scale up existing multivariate EVT models with the dimensionality. In the present work we bridge the gap between the practice of anomaly detection and multivariate EVT by proposing a method which is able to learn a sparse 'normal profile' of multivariate extremes and, as such, may be implemented to improve the accuracy of any usual anomaly detection algorithm.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ $i.i.d.$ random variables in $\mathbb{R}^d$ with joint (*resp.* marginal) distribution $\mathbf{F}$ (*resp.* F_j, $j = 1, \ldots, d$). Marginal standardization is a natural first step when studying the dependence structure in a multivariate setting. The choice of standard Pareto margins V^j (with $\mathbb{P}(V^j > x) = 1/x,\ x > 0$) is convenient, and justified by multivariate extreme value theory. One classical way to standardize is the probability integral transform, $T : X_i \mapsto V_i = ((1 - F_j(X_i^j))^{-1})_{1 \leq j \leq d}$, $i = 1, \ldots, n$. Since the marginal distributions F_j are unknown, we use their empirical counterpart $\hat{F}_j$, where $\hat{F}_j(x) = (1/n) \sum_{i=1}^n \mathbb{1}_{X_i^j \leq x}$. Denote by $\hat{T}$ the rank transformation thus obtained and by $\hat{V}_i = \hat{T}(X_i)$ the corresponding rank-transformed observations.

Now, for a given subset of features $\alpha \subset \{1, ..., d\}$ the goal is to measure the the likelihood to observe a large $\hat{\mathbf{V}}$ such that $\hat{V}^j$ is 'large' for all $j \in \alpha$, while the other $\hat{V}^j$'s ($j \notin \alpha$) are 'small'. That is, estimating $\Phi(\Omega_\alpha)$

In relation to Section 1, the appropriate way to give a meaning to 'large' (*resp.* 'small') among extremes is in 'radial' and 'directional' terms, that is, $\|\hat{\mathbf{V}}\| > r$ (for some high radial threshold r), and $\hat{V}^j / \|\hat{\mathbf{V}}\| > \epsilon$ (*resp.* $\leq \epsilon$) for some small directional tolerance parameter $\epsilon > 0$. Introduce the truncated ϵ-cones (see Fig. 2): $\mathcal{C}_\alpha^\epsilon = \big\{\mathbf{v} \geq 0,\ \|\mathbf{v}\|_\infty \geq 1,\ v_i > \epsilon \|\mathbf{v}\|_\infty \ \text{ for } i \in \alpha,\ \ v_i \leq \epsilon \|\mathbf{v}\|_\infty \ \text{ for } i \notin \alpha \ \big\}$, which defines a partition of $\mathbb{R}_+^d \setminus [0,1]^d$ for each fixed $\epsilon \geq 0$. This leads to estimates

$$\Phi_n^{\alpha,\epsilon} = (n/k)\hat{\mathbb{P}}_n\left((n/k)\mathcal{C}_\alpha^\epsilon\right), \tag{1}$$

where $\hat{\mathbb{P}}_n(.) = (1/n)\sum_{i=1}^n \delta_{\hat{\mathbf{V}}_i}(.)$ is the empirical probability distribution of the rank-transformed data and $k = k(n) \to \infty$ s.t. $k = o(n)$ as $n \to \infty$. The ratio n/k plays the role of a large radial threshold r. This estimator is justified by the result from multivariate EVT, $r\mathbb{P}(\mathbf{V} \in rA) \to \Phi(A)$ as $r \to \infty$.

This heuristic yields the following algorithm. The complexity is in $O(dn\log n + dn) = O(dn\log n)$, where the first term on the left-hand-side comes from computing the $\hat{F}_j(X_i^j)$ (Step 1) by sorting the data (*e.g.* merge sort). The second one comes from Step 2.

Algorithm 1. *(DAMEX)*
Input: *parameters* $\epsilon > 0, \quad k = k(n), \quad \Phi_{\min} \geq 0.$

1. *Standardize* via *marginal rank-transformation:* $\hat{V}_i := \hat{T}(X_i) = \big(1/(1 - \hat{F}_j(X_i^j))\big)_{j=1,\dots,d}$.
2. *Assign to each* $\hat{V}_i$ *the cone* $\mathcal{C}_\alpha^\epsilon$ *it belongs to.*
3. *Compute* $\Phi_n^{\alpha,\epsilon}$ *the estimate of the* α*-mass of* Φ *from (1).* $\to$ *yields: (small number of) cones with non-zero mass.*
4. *Set to* 0 *the* $\Phi_n^{\alpha,\epsilon}$ *below some small threshold* $\Phi_{\min} \geq 0$ *to eliminate cones with negligible mass*

Output: *(sparse) representation of the dependence structure*

$$\widehat{\mathcal{M}}(\alpha) := (\Phi_n^{\alpha,\epsilon})_{\alpha \subset \{1,\dots,d\}, \Phi_n^{\alpha,\epsilon} > \Phi_{\min}}$$

This algorithm can directly be applied to anomaly detection. The underlying assumption is that an observation is potentially abnormal if its 'direction' (after a standardization of each marginal) is special regarding to the other extreme observations. In other words, if it does not belong to the (sparse) support of extremes. The degree of 'abnormality' of new observation $\mathbf{x}$ such that $\hat{T}(\mathbf{x}) \in \mathcal{C}_\alpha^\epsilon$ should be related both to $\Phi_n^{\alpha,\epsilon}$ and the uniform norm $\|\hat{T}(\mathbf{x})\|_\infty$ (angular and radial components). As a matter of fact, in the transformed space - namely the space of the $\hat{V}_i$'s - the asymptotic mass decreases as the inverse of the norm. Consider the '*directional tail region*' induced by $\mathbf{x} \in \mathcal{C}_\alpha^\epsilon$, $A_\mathbf{x} = \{\mathbf{y} : T(\mathbf{y}) \in \mathcal{C}_\alpha^\epsilon,\ \|T(\mathbf{y})\|_\infty \geq \|T(\mathbf{x})\|_\infty\}$. Then, if $\|T(\mathbf{x})\|_\infty$ is large enough, it can be shown that $\mathbb{P}(\mathbf{X} \in A_\mathbf{x}) \simeq \|\hat{T}(\mathbf{x})\|_\infty^{-1}\Phi_n^{\alpha,\epsilon}$. We then set the scoring function $s_n(\mathbf{x}) := (1/\|\hat{T}(\mathbf{x})\|_\infty)\sum_\alpha \Phi_n^{\alpha,\epsilon}\mathbb{1}_{\hat{T}(\mathbf{x}) \in \mathcal{C}_\alpha^\epsilon}$, which is thus an empirical version of $\mathbb{P}(\mathbf{X} \in A_\mathbf{x})$: the smaller $s_n(\mathbf{x})$, the more abnormal the point $\mathbf{x}$ should be considered.

3 Theoretical grounds

The theorical results stated in this work build on the work of [9], where non-asymptotic bounds related to the statistical performance of a non-parametric estimator of the *stable tail dependence function* (STDF), another functional measure of the dependence structure of extremes, are established. However, even in the case of a sparse angular measure, the support of the STDF would not be so, since the latter functional is an integrated version of the former. Also, in many applications, it is more convenient to work with the angular measure. Indeed, it provides direct information about the probable 'directions' of extremes. To the best of our knowledge, non-parametric estimation of the angular measure has only been treated in the two dimensional case, in [6] and [5], in an asymptotic framework.

Under non-restrictive assumptions standard in EVT (existence of the angular measure and continuous marginal c.d.f.), we obtain a non-asymptotic bound of the form

$$\sup_{\emptyset \neq \alpha \subset \{1,\,\dots,\,d\}} |\widehat{\mathcal{M}}(\alpha) - \mathcal{M}(\alpha)| \;\leq\; Cd\left(\sqrt{\frac{1}{\epsilon k}\log\frac{d}{\delta}} + Md\epsilon\right) + \text{bias}(\epsilon, k, n),$$

with probability greater than $1 - \delta$, where $k = k(n) \to \infty$ with $k(n) = o(n)$ can be interpreted as the number of data considered as extreme. The bias term goes to zero as $n \to \infty$, for each fixed ϵ.

References

[1] D.A. Clifton, L. Tarassenko, N. McGrogan, D. King, S. King, and P. Anuzis. Bayesian extreme value statistics for novelty detection in gas-turbine engines. In *Aerospace Conference, 2008 IEEE*, pages 1–11, 2008.

[2] David Andrew Clifton, Samuel Hugueny, and Lionel Tarassenko. Novelty detection with multivariate extreme value statistics. *Journal of signal processing systems*, 65(3):371–389, 2011.

[3] SG Coles and JA Tawn. Modeling extreme multivariate events. *JR Statist. Soc. B*, 53:377–392, 1991.

[4] D. Cooley, R.A. Davis, and P. Naveau. The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117, 2010.

[5] J. H. J. Einmahl and J. Segers. Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, pages 2953–2989, 2009.

[6] John HJ Einmahl, Laurens de Haan, and Vladimir I Piterbarg. Nonparametric estimation of the spectral measure of an extreme value distribution. *Annals of Statistics*, pages 1401–1423, 2001.

[7] Barbel Finkenstadt and Holger Rootzén. *Extreme values in finance, telecommunications, and the environment*. CRC Press, 2003.

[8] Anne-Laure Fougères, John P Nolan, and Holger Rootzén. Models for dependent extremes using stable mixtures. *Scandinavian Journal of Statistics*, 36(1):42–59, 2009.

[9] N. Goix, A. Sabourin, and S. Clémençon. Learning the dependence structure of rare events: a non-asymptotic study. In *Proceedings of the 28th Conference on Learning Theory*, 2015.

[10] H.J. Lee and S.J. Roberts. On-line novelty detection using the kalman filter and extreme value theory. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4, 2008.

[11] S.J. Roberts. Novelty detection using extreme value statistics. *Vision, Image and Signal Processing, IEE Proceedings -*, 146(3):124–129, Jun 1999.

[12] S.J. Roberts. Extreme value statistics for novelty detection in biomedical signal processing. In *Advances in Medical Signal and Information Processing, 2000. First International Conference on (IEE Conf. Publ. No. 476)*, pages 166–172, 2000.

[13] Anne Sabourin and Philippe Naveau. Bayesian dirichlet mixture model for multivariate extremes: A re-parametrization. *Computational Statistics & Data Analysis*, 2012.

[14] RL Smith. Statistics of extremes, with applications in environment, insurance and finance, chap 1. *Statistical analysis of extreme values: with applications to insurance, finance, hydrology, and other fields. Birkhäuser, Basel*, 2003.

[15] A.G. Stephenson. High-dimensional parametric modelling of multivariate extreme events. *Australian & New Zealand Journal of Statistics*, 51(1):77–88, 2009.