Journal of Multivariate Analysis 161 (2017) 12–31

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Sparse representation of multivariate extremes with applications to anomaly detection

Nicolas Goix, Anne Sabourin, Stephan Clémençon *

LTCI, Télécom ParisTech, Université Paris-Saclay 46, rue Barrault, 75013, Paris, France

ARTICLE INFO

Article history:
Received 28 January 2016
Available online 14 July 2017

AMS subject classifications:
62H12
62G32

Keywords:
Anomaly detection
Dimensionality reduction
Multivariate extremes
VC theory

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 extreme dependence in the framework of multivariate Extreme Value Theory (EVT) is the angular measure, which provides direct information about the probable "directions" of extremes, i.e., 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. The present paper proposes a novel methodology aiming at exhibiting a particular kind of sparsity within the dependence structure of extremes. This is achieved by estimating the amount of mass spread by the angular measure on representative sets of directions corresponding to specific sub-cones of $\mathbb{R}_+^d$. This dimension reduction technique paves the way towards scaling up existing multivariate EVT methods. Beyond a non-asymptotic study providing a theoretical 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 behaviors, to be confronted with new (possibly abnormal) extreme observations. Illustrative experimental results provide strong empirical evidence of the relevance of our approach.

1. Introduction

1.1. Context: multivariate extreme values in large dimension

Extreme Value Theory (EVT in abbreviated form) provides a theoretical basis for modeling the tails of probability distributions. In many applied fields where rare events may have a disastrous impact, such as finance, insurance, climate, environmental risk management, network monitoring [23,43] or anomaly detection [8,31], the information carried by extremes is crucial. In a multivariate context, the dependence structure of the joint tail is of particular interest, as it gives access 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 simultaneously large while the others remain small, which is a valuable piece of information for multi-factor risk assessment or detection of anomalies among other non abnormal extreme data.

In a multivariate "Peak-Over-Threshold" setting, realizations of a d-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ are observed and the goal pursued is to learn the conditional distribution of excesses, $\mathbf{Y} \mid \|\mathbf{Y}\| \geq r$, above some large threshold

* Corresponding author.
E-mail address: stephan.clemencon@telecom-paristech.fr (S. Clémençon).

http://dx.doi.org/10.1016/j.jmva.2017.06.010
0047-259X/

$r > 0$. 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 (see Section 2), $C\Phi(A) \simeq \Pr(\|\mathbf{Y}\|^{-1}\mathbf{Y} \in A \mid \|\mathbf{Y}\| > r)$, 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 $\{\max_{1\le i\le k} x_i = 1 : x_1 > 0, \ldots, x_k > 0, x_{k+1} = \cdots = x_d = 0\}$ indicates that the components $Y_1, \ldots, Y_k$ may be large simultaneously, while the others are small. An extensive exposition of this multivariate extreme setting may be found, e.g., in [4,35].

Parametric or semi-parametric modeling and estimation of the structure of multivariate extremes is relatively well documented in the statistical literature; see, e.g., [11,12,25,38] and the references therein. In a nonparametric setting, there is also an abundant literature concerning consistency and asymptotic normality of estimators of functionals characterizing the extreme dependence structure, e.g., extreme-value copulas or the stable tail dependence function (STDF); see [13,16,18, 21,26]. In many applications, it is nevertheless more convenient to work with the angular measure itself, as the latter gives more direct information on the dependence structure and is able to reflect structural simplifying properties (e.g., sparsity as detailed below) which would not appear in copulas or in the STDF. However, nonparametric modeling of the angular measure faces major difficulties, stemming from the potentially complex structure of the latter, especially in a high-dimensional setting. Further, from a theoretical point of view, nonparametric estimation of the angular measure has mainly been studied in the bivariate case, in [17] and [20], in an asymptotic framework (high-dimensional cases being discussed in [18]).

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 $d \le 10$). 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., [45] or [38]. 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.

In this paper, we consider situations where phenomenon (I) below occurs, and possibly phenomenon (II) as well. As illustrated by the examples in Section 5, these can be encountered in several applications.

(I) 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 (the adjective small is relative to the total number of groups 2^d).
(II) 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.

In order to gain more insight into the nature of the conditions stated above and understand better why one may reasonably expect that one and/or the other are fulfilled in common practical situations, consider the three following stylized examples: (i) there is a fully dependent asymptotic structure: one feature is large if and only if all the features are large; (ii) there is an asymptotic independence structure: a feature can only be extreme on its own; (iii) there is a full dependence structure pattern: given the fact that a feature j is large, the group of features being extreme can be any group of features containing j. In situation (i), only the central hyper-cube corresponding to feature group $\{1, \ldots, d\}$ (i.e., the cartesian product $(1, \infty]^d$ for an appropriate standardization of the data) has non-zero mass. Thus phenomenon (I) is observed, but not phenomenon (II) since this group is not small (the corresponding hyper-cube does not have a small dimension). In situation (ii), only the d hyper-cubes corresponding to the singletons $\{1\}, \ldots, \{d\}$ have non-zero mass. As $d \ll 2^d$, phenomenon (I) occurs. As $1 \ll d$, so does phenomenon (II). Finally, in situation (iii), any hyper-cube has non-zero mass, so that neither phenomena occurs. In many real-word cases, one expects to may be far from situation (iii), namely to be in the range of phenomenon (I), allowing for a sparse description of the extremal dependence structure. Incidentally, we point out that, of course, this notion of sparsity in the extremes is by no means the sole that can be considered but it offers obvious advantages regarding interpretability.

The main purpose of this paper is to introduce a data-driven methodology for detecting this specific sparse structure when it occurs and recovering the corresponding hyper-cubes, so as to possibly reduce the dimensionality of the problem and thus to learn an interpretable representation of extreme behaviors. In case hypothesis (II) is not fulfilled, such a sparse profile can still be learned, but loses 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 hyper-cube.

The theoretical results stated in this paper build on the work of [27], where non-asymptotic bounds related to the statistical performance of a nonparametric estimator of the 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; see Eq. (2.2), Section 2. Also, in many applications, it is more convenient to work with the angular measure. Indeed, it provides direct information about the probable directions of